Verteilte Anonymisierung von vertikal partitionierten Daten

Diplomarbeit

zur Erlangung des akademischen Grades
Diplominformatiker(in)

Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät II
Institut für Informatik

eingereicht von: Jan Hendrik Nielsen

Gutachter(innen): Herr Prof. Johann-Christoph Freytag, Ph. D.
Herr Prof. Dr. Wolfgang Reisig

eingereicht am: 22. März 2013

ZUSAMMENFASSUNG

Das Erheben und Verarbeiten von sensiblen, personenbezogenen Informationen zur statistischen Auswertung ist im medizinischen Umfeld unerlässlich. Aufgrund der Vertraulichkeit der Daten kann eine Veröffentlichung ausschließlich anonymisiert erfolgen.

Die De-Identifikation der Daten durch das Entfernen direkt identifizierender Attribute wie dem Namen reicht jedoch nicht aus um die Privatsphäre eines Individuums zu schützen.

Moderne Konzepte zum Schutz der Privatsphäre schaffen die Voraussetzung zur Veröffentlichung der Daten unter Einhaltung strenger Datenschutzrichtlinien. Das Konzept der k-Anonymisierung ermöglicht eine Veröffentlichung der Daten ohne deren Semantik zu verändern. Zu diesem Zweck folgt das Prinzip der k-Anonymisierung syntaktischen Vorgaben bezüglich der Form der Veröffentlichung. Die t-Closeness stellt eine vielbeachtete Weiterentwicklung der k-Anonymisierung dar. Sie bezieht die Semantik der Daten in die Anonymisierung ein.

Diese Konzepte wurden jedoch nicht für die Anonymisierung räumlich getrennter Daten entwickelt. Dieses Problem entsteht durch die zunehmende Dezentralisierung von Daten. Vielfach erheben verschiedene Institutionen Daten unterschiedlicher Semantik über die gleiche Person. Diese vertikale Partitionierung der Daten stellt neue Anforderungen an Verfahren zum Schutz der Privatsphäre.

Während Methoden der dezentralen Anonymisierung mithilfe der k-Anonymisierung existieren, ist dies für das Konzept der t-Closeness nicht der Fall. Die vorliegende Arbeit wird diese Lücke schließen und Anforderungen analysieren, unter denen eine verteilte Datenanonymisierung über vertikal partitionierten Daten mittels des Konzepts der t-Closeness möglich ist. Auf dieser Grundlage wird ein kryptografisches Protokoll zur verteilten Datenanonymisierung mithilfe der t-Closeness konzeptionell entwickelt.

INHALTSVERZEICHNIS

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

ALGORITHMENVERZEICHNIS

ABKÜRZUNGSVERZEICHNIS

SMC	Secure-Multiparty-Computation
HBC	Honest-But-Curious
TTP	Trusted-Third-Party
DBMS	Datenbank-Management-System
DPP_2GA	Distributed Privacy-Preserving two-Party Generic Anonymizer
SSI	Secure Set Intersection
FtEDPP_2GA	Fragmenting t-Closeness-Enhanced DPP_2GA
EMD	Earth-Movers-Distanz
KL-Distanz	Kullback-Leibler-Distanz
DGH	Domain-Generalization-Hierarchy
VGH	Value-Generalization-Hierarchy
MSA	Multiple-Sensitive-Attributes

CRP	C-reaktives Protein
L-WBC	Leukozyten der Gehirn-Rückenmarks-Flüssigkeit
MH-Distanz	Manhattan-Distanz
ID	Identifikator
QID	Quasi-Identifikator
BID	Bucket-Identifikator
SA	Sensibles Attribut
PPDM	Privacy-Preserving Data Mining
DM	Data Mining
PPDP	Privacy-Preserving Data Publishing
FDG	Full-Domain-Generalisierung
DF-Algorithmus	Datafly-Algorithmus
HIPAA	Health Insurance Portability and Accountability Act
BDSG	Bundesdatenschutzgesetz
BStatG	Bundesstatistikgesetz
PbD	Privacy by Design

1 EINLEITUNG

1.1 MOTIVATION

> „You have zero privacy anyway. Get over it."
>
> — Scott McNealy (ehemaliger Vorstand von SUN Microsystems) [Spr99]

Dieses Zitat sorgte nach seiner Äußerung im Jahre 1999 für Aufsehen. Der Vorstand eines großen Softwareunternehmens sprach seinen Kunden das Recht auf Privatsphäre ab. Um das Zitat genauer einordnen zu können, müssen wir uns zunächst mit der Frage beschäftigen, was unter *Privatsphäre* zu verstehen ist. Alan Westin, emeritierter Professor für öffentliches Recht der Columbia-Universität, definiert den Begriff wie folgt: „Privacy is the right of individuals to determine for themselves when, how and to what extent information about them is communicated to others." [Agr+02, S. 1]. Diese Definition offenbart die subjektive Wahrnehmung des Begriffs der Privatsphäre. Gleichzeitig assoziiert sie Privatsphäre mit dem Begriff der *Information*. Information über ein Individuum wird aus *Daten* über eben dieses gewonnen. Wegen des individuellen Charakters der Daten bezeichnen wir sie als *personenbezogene Daten*. Durch sie wird ein Individuum unterscheidbar von anderen Individuen [Koh05, S. 32f]. Der Begriff der personenbezogenen Daten ist im Bundesdatenschutzgesetz (BDSG) §3 Absatz 1 eigenständig definiert [Deu09]. Das BDSG attestiert ihnen eine besondere Schutzwürdigkeit. Die Art und Weise des Schutzes ist hingegen nicht genauer erfasst. Er kann sich sowohl auf die Korrektheit der Daten, als auch auf deren Vertraulichkeit beziehen [Wei06]. Erschwerend kommt hinzu, dass die Definition von Privatsphäre kontextabhängig betrachtet werden muss, wird durch eine häufig verwendete Definition von Latanya Sweeney deutlich: „Privacy reflects the ability of a person, organization, government, or entity to control its own space, where the concept of space (or 'privacy space') takes on different contexts" [Swe03]. In der Folge wandelt sich der subjektive Begriff der Privatsphäre in dem Kontext in dem er verwendet wird. Die Wandlung des Kontextes wird an folgendem Beispiel deutlich: Während es auf der einen Seite unerwünscht sein kann Adressdaten in einem Telefonbuch zu veröffentlichen, so ist es andererseits nicht zu verhindern, dass diese der örtlichen Meldebehörde mitzuteilen sind. In der Folge existieren Sammlungen personenbezogener Daten, auf deren Veröffentlichung das betroffene Individuum keinen Einfluss hat. Beispielsweise kann die Veröffentlichung der Daten, wie

im Fall der Krebsregister der Bundesrepublik Deutschland[1], gesetzlich vorgeschrieben sein. Diese überregionale Erhebung von Krankheitsverläufen bei der Behandlung von Krebspatienten und deren Veröffentlichung soll es ermöglichen, bessere Behandlungsmöglichkeiten zu entwickeln. Es ist unmittelbar ersichtlich, dass die Veröffentlichung von medizinischen Daten ein sehr sensibles Thema darstellt. Die öffentliche Assoziation von Krankheiten und Individuen könnte fatale Folgen für die Betroffenen haben. Eine Studie von Linowes et al. verdeutlichte, dass 35 % der 87 größten, weltweit operierenden Unternehmen sich in ihren Personalentscheidungen durch Krankheitsdiagnosen beeinflussen ließen [LS90]. Aus diesem Grund muss die Identität eines Individuums bei der Veröffentlichung derartiger Daten geschützt bleiben, darf also nicht nachvollziehbar sein.

Der Schutz der Privatsphäre bei statistischen Veröffentlichungen ist ein wohlbekanntes und umfassend erforschtes Thema [Dwo11]. Erste Abhandlungen der amerikanischen Volkszählungsbehörde „Bureau of Census“ zum Thema Datenschutz von Veröffentlichungen stammen aus dem Jahr 1929 [Dal77]. Statistische Daten bestehen im Allgemeinen nicht aus den Originalwerten der Erhebung. Vielmehr werden sie in aufgearbeiteter Form veröffentlicht. Beispielsweise werden Daten aggregiert oder unter Beibehaltung der statistischen Eigenschaften derart verändert, dass die entstehenden *Makrodaten* keine Rückschlüsse auf Individuen zulassen [Karo8, S. 18f]. In diesem Zusammenhang sprechen wir von einer *Anonymisierung* der Daten.

Anonymisierte Makrodaten bieten nur einen begrenzten Nutzen. Gerade im Bereich der Medizin ist es notwendig Zugriff auf die Originaldaten zu erhalten. Aufgrund der zuvor genannten Sensibilität medizinischer Daten sieht das Bundesstatistikgesetz die Weitergabe der Originaldaten an Forschungseinrichtungen in anonymisierter Form vor [Karo8, S. 18f]. Ursprünglich wurden aus diesen als *Mikrodaten* bezeichneten Daten sämtliche direkt identifizierenden Attribute entfernt. Hierbei kann es sich beispielsweise um den Namen oder eine eindeutige Nummer handeln. Ein solcher Vorgang wird als *De-Identifikation* bezeichnet [Swe97b, S. 7]. Dass diese Maßnahme zu kurz greift, wurde in einer viel beachteten Studie von Latanya Sweeney nachgewiesen [Sweoo]. Sweeney konnte nachweisen, dass die Kombination mehrerer nicht-identifizierender Attribute einer Person einzigartig sein kann. Ist eine Kombination dieser Art ebenfalls in einer anderen Datenquelle vorhanden, welche nicht anonymisiert wurde, so kann die Verknüpfung beider Datenquellen zur *Re-Identifikation* führen. Konkret konnten anhand der Attribute Geschlecht, Alter und Postleitzahl und deren Abgleich mit öffentlich verfügbaren Wählerlisten ca. 87 % der de-identifizierten Personen re-identifiziert werden. Der Vorgang der Verknüpfung zweier Datenquellen wird *Linking* genannt. Die Verfügbarkeit von für Linking

1 http://www.gekid.de/

relevanten, nicht anonymisierten Datenquellen ist kein theoretisches Problem. Durch neue Technologien erschließen sich bis dato ungeahnte Möglichkeiten Daten zu erfassen und dauerhaft zu sichern. Algorithmen des *Data Minings* ermöglichen das Extrahieren nutzbringender Information, aus sehr großen Datenbeständen in kurzer Zeit [HKP12, S. 31]. Laut Sweeney kann ein Trend im Sammeln von personenbezogenen Daten beobachtet werden [Swe01, S. 41].

Rekapitulieren wir die bisherigen Erkenntnisse, so ergibt sich durch die hohe Verfügbarkeit personenbezogener Daten eine konkrete Gefahr der Re-Identifikation durch Linking. Die Anonymisierung von Mikrodaten durch De-Identifikation würde sich als nutzlos erweisen. In der Folge besäßen veröffentlichte sensible Daten keinen Schutz. Es träfe die Feststellung der Eingangs bemerkten Aussage von Scott McNealy zu: Es ließe sich kein Schutz sensibler Daten erreichen, jedes Individuum besäße „zero privacy“.

Um die Schwächen der De-Identifikation zu lösen, wurde das Konzept der k-Anonymisierung entwickelt [Swe02b]. Es erweitert das Prinzip der De-Identifikation um diejenigen Attribute, welche für Linking infrage kommen. Diese werden im Gegensatz zu den direkt identifizierenden Attributen nicht entfernt. Vielmehr werden ihre Werte derart verändert, dass sich stets k Individuen bezüglich dieser Werte nicht unterscheiden. Die Zuordnung eines sensiblen Attributs ist somit nur noch zu k Individuen möglich. Das Konzept der k-Anonymisierung begründete den Anfang einer Reihe von erweiterten Konzepten des Datenschutzes durch Anonymisierung der Daten. Die bekanntesten Konzepte, ℓ-Diversity und t-Closeness, werden uns im Laufe dieser Diplomarbeit beschäftigen.

Durch k-Anonymisierung oder seine Erweiterungen können Veröffentlichungen von Mikrodaten sinnvoll geschützt werden. Jedoch erscheint im Kontext medizinischer Studien ein weiteres Problem. Medizinische Studien werden häufig von mehr als einer Institution durchgeführt [Sch09]. Studien dieser Art werden als *multizentrische Studien* bezeichnet. Sie sind aufgrund ihrer erhöhten Aussagekraft gegenüber monozentrischen Studien vorzuziehen [Sch09]. Weiterhin werden sie vom Bundesministerium für Bildung und Forschung gefördert[2]. In diesem Zusammenhang sprechen wir von einem *verteilten* oder *dezentralen* System. Durch die räumliche Trennung der an der Studie beteiligten Institutionen, ergeben sich mehrere Probleme. Zum einen werden zwangsläufig die erhobenen Daten getrennt. Zum anderen können die beteiligten Institutionen verschiedenen Jurisdiktionen unterstellt sein, die unterschiedliche Vorgaben an den Datenschutz aufweisen. Letzteres führt dazu, dass der Austausch von Originaldaten zwischen den Institutionen unmöglich wird. Der Austausch hat in anonymisierter Form zu erfolgen. Die Konzepte k-Anonymisierung,

2 http://www.gesundheitsforschung-bmbf.de/de/308.php

ℓ-Diversity und t-Closeness wurden jedoch nicht im Hinblick auf ein verteiltes Szenario entwickelt. Es besteht daher ein Bedarf an Mechanismen zur Anonymisierung verteilt erhobener Daten und deren Zusammenführung. Die Anforderungen an ein solches System sollen uns nachfolgend beschäftigen.

1.2 ZIELSETZUNG DER ARBEIT

Die vorliegende Diplomarbeit widmet sich nun der konzeptionellen Erschließung vorhandener Methoden des Datenschutzes durch Datenanonymisierung. Dabei betrachten wir Daten im Folgenden als Realisierung eines relationalen Datenbankschemas, dessen Darstellung als Tabelle erfolgt. Der Schwerpunkt liegt auf den Konzepten der k-Anonymisierung, ℓ-Diversity und der t-Closeness, über deren Stärken und Schwächen die Arbeit einen Überblick geben wird. Das Fundament der Datenanonymisierung bildet die k-Anonymisierung. Um Schwachstellen der k-Anonymisierung zu korrigieren, wurde das Konzept der ℓ-Diversity entwickelt. Die Intention der ℓ-Diversity fand durch das Konzept der t-Closeness eine weitere Verfeinerung. Diese drei Konzepte bilden den Fortschritt der Datenanonymisierung ab und sind als Ausgangspunkt für die Entwicklung weiterer Methoden des Datenschutzes von besonderer Bedeutung.

Es existieren Ansätze, welche die Konzepte der k-Anonymisierung und ℓ-Diversity in einem verteilten Szenario realisieren. An dieser Stelle ist es sinnvoll, die Art der Verteilung der Daten genauer zu klassifizieren. Grundsätzlich wird zwischen *horizontaler* und *vertikaler* Verteilung unterschieden. In diesem Zusammenhang wird oftmals von einer horizontalen oder vertikalen *Partitionierung* der Daten gesprochen. Die vertikale Partitionierung von Daten beschreibt nach Buchmann die Teilung von „Daten *unterschiedlicher Semantik* über die *gleichen Individuen*“ auf mehrere Tabellen [Buc10]. Eine horizontale Partitionierung der Daten ist gegensätzlich zu verstehen. Hier sind „Daten *gleicher Semantik* über *unterschiedliche Individuen*“ auf mehrere Tabellen verteilt [Buc10].

Kernpunkt der Diplomarbeit ist es, Anforderungen zu analysieren, unter denen eine verteilte Datenanonymisierung über vertikal partitionierten Daten mittels des Konzepts der t-Closeness möglich ist. Anschließend wird ein Protokoll zur verteilten Datenanonymisierung mithilfe der t-Closeness entwickelt. Als Anwendungsszenario wird von einer multizentrischen medizinischen Studie ausgegangen. An der Studie sind zwei Institutionen beteiligt. Die Institutionen verfügen jeweils über denselben Patientenstamm, d. h. jedes Individuum der Datenbasis einer Institution findet sich in der Datenbasis der anderen Institution. Von diesem erhebt jede Institution eine unbestimmt Anzahl an Attributen. Die Menge der Attribute besteht jeweils

aus einem sensiblen Attribut, dessen Schutz durch die verteilte Anonymisierung gewährleistet werden soll, sowie einer beliebigen Anzahl an Attributen welche für Linking anfällig sind. Zusätzlich legen wir fest, dass die Attribute der beteiligten Institutionen unterschiedlich sein müssen.

Abschließend sei darauf hingewiesen, dass die Bedrohungen sensibler Daten in Datenbank-Management-Systemen vielfältiger Natur sind. Während schon der physische Zugriff auf die gespeicherten Daten als Verletzung der Privatsphäre gesehen werden kann, liegt die Betrachtung von Zugriffskontroll- und Auditing-Mechanismen außerhalb des Rahmens dieser Diplomarbeit. Wir gehen für den Rest der Diplomarbeit davon aus, dass ein geeigneter Schutz der Daten besteht.

1.3 AUFBAU DER ARBEIT

Die Diplomarbeit gliedert sich in sechs Kapitel, deren Schwerpunkte wir nachfolgend betrachten wollen.

In Kapitel 2 werden benötigte mathematische Grundlagen gelegt. Dies dient einerseits dem Verständnis des Lesenden und ermöglicht andererseits die Definition einer einheitlichen Syntax, welche wir auf die gesamte Arbeit übertragen werden.

Kapitel 3 erläutert dem Lesenden grundlegende Begriffe der Datenanonymisierung. Es vermittelt einen Überblick über Datenstrukturen, auf denen die Konzepte der Datenanonymisierung aufbauen. Dieses Kapitel wird außerdem die Verknüpfung der mathematischen Theorie mit der Anonymisierung von sensiblen Daten ermöglichen.

Eine Einführung in Methoden der Datenanonymisierung bietet Kapitel 4. Es werden drei Anonymisierungskonzepte genauer untersucht. Namentlich die k-Anonymisierung, ℓ-Diversity sowie die t-Closeness. In diesem Kapitel werden wir die Schwächen der Konzepte durch existierende Angriffe auf die Anonymisierung aufzeigen.

Kapitel 5 bildet den Hauptteil der Arbeit. In diesem werden die vorher erläuterten Konzepte bezüglich der Anwendbarkeit in einem verteilten Szenario untersucht. Der Fokus liegt auf den Konzepten der k-Anonymisierung sowie der t-Closeness.

Die Diskussion des entwickelten Konzepts erfolgt in Kapitel 6. Ferner werden an dieser Stelle weitere Konzepte des Datenschutzes diskutiert. Abschließend wird die Arbeit in einen größeren Forschungszusammenhang eingeordnet.

2 MATHEMATISCHE GRUNDLAGEN

Bevor wir uns den Algorithmen zum Schutz sensibler Daten zuwenden, müssen wir zunächst eine theoretische Grundlage schaffen. Durch sie werden wir über eine einheitliche Syntax verfügen. In diesem Kapitel werden wir uns daher vorerst der Terminologie zuwenden, welche zum weiteren Verständnis der Anonymisierungskonzepte notwendig ist.

2.1 EINFÜHRUNG DER VERWENDETEN MATHEMATISCHEN BEGRIFFE

Mathematische Grundlagen der Analysis werden für diese Arbeit vorausgesetzt. Ein zentraler Begriff dieser Arbeit ist jener der *Metrik*. Aufgrund seiner bedeutenden Stellung soll er an dieser Stelle formalisiert werden, bevor wir uns im weiteren Verlauf der Arbeit der deskriptiven Statistik zuwenden.

Metriken sind Funktionen, welche Elemente einer beliebigen Menge M in den Raum der positiven Reellen Zahlen $\mathbb{R}_+$ abbilden – ihnen also einen messbaren Wert zuordnen. Es werden uns Metriken begegnen, welche den Grad einer Datenanonymisierung widerspiegeln, Metriken deren Zweck es ist die Güte einer Anonymisierung zu messen sowie Metriken, welche dazu benutzt werden, um eine „optimale“[1] Anonymisierung zu erzielen. Metriken messen den *Abstand* zwischen Elementen, der ihnen durch ihre natürliche Ordnung innewohnt. Dieser Abstand wird durch eine Funktion $d(\cdot\|\cdot)$, die sogenannte *Abstandsfunktion* oder das *Abstandsmaß*, definiert[2]. Sie hat nach Bronstein et al. [Bro+01] die folgenden Eigenschaften:

Definition 1 (Metrik)**.** *Sei M eine Menge, x, y, z beliebige Elemente aus M und d eine Abbildung, für die gelte:*

$$d : M \times M \longrightarrow \mathbb{R}_+ := \{x \in \mathbb{R} | x \geq 0\}$$

genannt Metrik *auf M, welche folgenden Axiomen genügt:*

$$d(x\|y) \geq 0 \text{ und } d(x\|y) = 0 \iff x = y \qquad \text{(Nichtnegativität)}, \qquad \text{(M1)}$$

$$d(x\|y) = d(y\|x) \qquad \text{(Symmetrie)}, \qquad \text{(M2)}$$

$$d(x\|z) \leq d(x\|y) + d(y\|z) \qquad \text{(Dreiecks-Ungleichung)}. \qquad \text{(M3)}$$

1 An dieser Stelle kann nicht darauf eingegangen werden, was *optimal* bedeutet, die Interpretation ist kontextabhängig.

2 Diese „Punktnotation“ wird im Folgenden der Übersichtlichkeit halber Verwendung finden, sobald eine konkrete Variable für das Verständnis der Funktion nicht notwendig ist.

$d(x\|y)$ heißt der Abstand von x und y. Die mit einer Metrik versehene Menge M heißt metrischer Raum.

Abseits dieses fundamentalen Abstandsbegriffs bedient sich diese Diplomarbeit der *deskriptiven* oder *beschreibenden Statistik*. Die in diesem Teilgebiet verwendeten Termini werden in den verschiedenen Sprachräumen unterschiedlich notiert und auch im deutschen Sprachraum ist die Notation uneinheitlich. Im Folgenden werden wir für den Kontext dieser Arbeit diesbezüglich Klarheit schaffen und dabei Gegebenheiten betrachten, die durch das Anwendungsgebiet der Datenanonymisierung entstehen. Die nachfolgenden Notationen sind Kohn [Koh05] entnommen, soweit dies nicht anders angegeben wird.

2.1.1 *Grundbegriffe der deskriptiven Statistik*

Die deskriptive Statistik beschäftigt sich mit der Beschreibung und Darstellung von *statistischen Daten* [Fah+07, S. 11]. Statistische Daten – oder kurz *Daten* – bezeichnen ermittelte *Informationen*, welche sich *aggregieren* oder *quantifizieren* lassen. Diese Daten werden aus *Objekten* gewonnen, durch deren Beobachtung Aussagen über einen zu untersuchenden Sachverhalt getroffen werden können. Diese Objekte heißen *Merkmalsträger* oder *Untersuchungseinheiten*. Im Kontext einer statistischen Erhebung kann aus vielerlei Gründen nicht immer auf alle theoretisch verfügbaren Merkmalsträger zurückgegriffen werden. Häufig wird daher ein Ausschnitt der prinzipiell verfügbaren Information betrachtet. Die Menge aller Merkmalsträger wird in ihrer Gesamtheit als *Population* oder *Grundgesamtheit* bezeichnet. Sie wird mit dem griechischen Ω gekennzeichnet. Die in ihr enthaltenen Merkmalsträger werden in der Regel als ω_i dargestellt. Der Laufindex i bezeichnet ihre Position innerhalb der Grundgesamtheit, ohne jedoch eine Wertung zu implizieren. Eine Teilmenge der Population wird als *Stichprobe* bezeichnet und bildet zumeist die Basis für statistische Untersuchungen. Die Stichprobe wird mit $\mathcal{X}$ bezeichnet. Der *Stichprobenumfang*, also die Anzahl der beobachteten Werte, wird gemeinhin mit der natürlichen Zahl n notiert. Statistische Beobachtungen beziehen sich nicht auf die Gesamtheit eines Merkmalsträgers. Es werden nur für den Kontext der Erhebung relevante und interessante Größen – die sogenannten *Merkmale* – bestimmt. Sie werden in der Regel durch einen großen lateinischen Buchstaben gekennzeichnet, zumeist X oder Y. Die Werte, die ein Merkmal annehmen kann, werden als *Merkmalsausprägungen* bezeichnet. Die Ausprägungen eines Merkmals X seien wie folgt bezeichnet:

$$\mathcal{A}_X = \{x_1, x_2, \ldots, x_m\} \; \forall m \in \mathbb{N} \text{ mit } m \leq n \tag{2.1}$$

Im Folgenden werden wir auf die explizite Notation des Wertebereichs der Indizes verzichten und annehmen, dass es sich um natürliche Zahlen handele, welche durch die Werte 1 sowie $m \leq n$ begrenzt seien. Analog bezeichnen wir Werte eines Merkmals Y mit $\mathcal{A}_Y = \{y_1, y_2, \ldots, y_m\}$. Der Index m wird in aller Regel von n verschieden und sogar kleiner sein, da es mehr Beobachtungen als mögliche oder interessierende Werte gibt. Dies ist selbstverständlich nicht immer der Fall, wir wollen diese Betrachtungsweise jedoch für den Rest der Arbeit beibehalten. Um die beobachteten Merkmalsausprägungen mit den möglichen Ausprägungen in Relation setzen zu können, definieren wir die Funktion $X : \mathcal{X} \longrightarrow \mathcal{A}_X$:

Definition 2 (Beobachtungswert)**.** *Die beobachtete Ausprägung eines Merkmals X bei einem Merkmalsträger ω_i wird* Beobachtungswert *genannt und durch x_i dargestellt. Wir schreiben $X(\omega_i) = x_i$ mit $x_i \in \mathcal{A}_X$*

Eine Stichprobe deckt nicht zwangsläufig alle möglichen Merkmalsausprägungen ab, sondern nur eine Teilmenge dieser: Die Menge der beobachteten Werte. Zusätzlich können Werte mehrfach auftreten. Die Beobachtungen der Ausprägungen eines Merkmals in einer Stichprobe werden als *Urliste* bezeichnet.

2.1.2 *Skalen und Typen von Merkmalen*

Im Rahmen dieser Arbeit betrachten wir nur Merkmale, welche eine endliche Anzahl an Werten annehmen können. Diese Art von Merkmalen nennt man *diskrete Merkmale*. Je nachdem in welcher Weise diese Werte voneinander unterschieden werden können, sprechen wir von *qualitativen*, *komparativen* und *quantitativen* Merkmalen. Um einem Wert eine Ausprägung zuordnen zu können, bedarf es einer *Skala*, welche sämtliche möglichen Ausprägungen abdeckt. Laut Cramer et al. ist eine Skala „eine Vorschrift, die jeder statistischen Einheit der Stichprobe einen Beobachtungswert zuordnet. Dieser Wert gibt die Ausprägung des jeweils interessierenden Merkmals an.“ [CKo8, S. 5]. Skalen werden durch ihr *Niveau* unterschieden. Die Skala mit dem niedrigsten Niveau ist die *Nominalskala*. Werte dieser Skala können nur nach den Kriterien „gleich“ und „verschieden“ voneinander abgegrenzt werden [Koho5, S. 13].

Merkmale deren Ausprägungen ein nominales Skaleniveau aufweisen werden *Qualitative Merkmale* genannt [Koho5, S. 13].

Definition 3 (Qualitatives Merkmal)**.** *Qualitative Merkmale lassen sich durch ihre Namen unterscheiden, besitzen jedoch keine Rangfolge. Sie können nur durch die Relationen „gleich“ und „verschieden“ voneinander abgegrenzt werden.*

Beispiel 1. Ein typisches Beispiel für ein qualitatives Merkmal ist das Geschlecht einer Person. Für gewöhnlich wird nur zwischen männlich und weiblich unterschieden. Dem jeweiligen Geschlecht wird keine hervorgehobene Bedeutung beigemessen, wodurch keine Ausprägung gegenüber der anderen bevorzugt wird.

Das nächsthöhere Skalenniveau wird von der *Ordinalskala* bedient. Werte, welche sich auf der Ordinalskala abbilden lassen, unterliegen einer Rangfolge. Die Abstände dieser Rangfolge lassen sich jedoch nicht interpretieren. Merkmale, deren Ausprägungen sich ordinal abbilden lassen, bezeichnet man als *komparative Merkmale*.

Definition 4 (Komparatives Merkmal)**.** *Komparative Merkmale besitzen im Vergleich zu qualitativen Merkmalen zusätzlich eine natürliche Rangfolge, welche sich jedoch nicht durch einen Abstand charakterisieren lässt.*

Medizinische Messwerte fallen häufig in diese Kategorie.

Beispiel 2. Beispielhaft hierfür ist die Kaliumkonzentration im Körper eines erwachsenen Menschen. Kalium ist ein Mineral, dessen Konzentration für viele physiologische Prozesse des Körpers wichtig ist. Für Erwachsene gilt ein Richtwert von $3,6$ bis $5,0$ $^{mmol}/_{l}$ [Hero8, S. 869]. Demnach gelten Werte innerhalb dieser Grenzen als „normal“, Werte darunter als „zu niedrig“ und darüber liegende Werte als „zu hoch“. Nun würde ein Wert von $3,5$ als zu niedrig angesehen werden und ein Wert von $7,0$ als zu hoch. Es lässt sich jedoch nicht folgern, dass der doppelt so große Wert doppelt so schlecht ist, da die Auswirkungen grundverschieden sind. Wir sehen also, dass komparative Merkmale durchaus aus numerischen Werten, d. h. Zahlenwerten, gebildet werden können. Es ist jedoch wichtig darauf hinzuweisen, dass zwar eine Rangfolge der Werte möglich ist, nicht jedoch auf ihnen eine Metrik zu definieren und aus dieser Schlüsse zu ziehen.

Qualitative und Komparative Merkmale lassen sich nur in Kategorien oder Klassen einteilen [Koh05, S. 14]. Aus diesem Grund spricht man bei ihnen auch von *kategorialen Merkmalen*.

Definition 5 (Kategoriales Merkmal)**.** *Ein kategoriales Merkmal ist ein Merkmal, welches Definition 3 oder 4 genügt und somit qualitativ oder komparativ ist.*

Anders verhält es sich mit *quantitativen Merkmalen*. Diese stehen in einer messbaren Relation zueinander und lassen sich daher auf einer *metrischen Skala* abbilden [Koh05, S. 14].

Definition 6 (quantitatives Merkmal)**.** *Der Wertebereich der Merkmalsausprägungen eines quantitativen Merkmals entstammt einer Zahlenmenge M welche eine* lineare Ordnung *besitzt und auf der ferner eine Metrik definiert ist.*

Die Ordnungsrelation ist *trichotom*, d. h. eindeutig in Bezug auf die Relationen „gleich“ und „größer oder kleiner als“, sowie abgeschlossen bezüglich Addition und Multiplikation [Foro8, S. 20]. Somit bildet die Menge M zusammen mit der Abstandsfunktion $d(x\|y)$ für $x, y \in M$ einen metrischen Raum (M, d).

Im Rahmen dieser Arbeit wird es sich bei der Menge M zumeist um die natürlichen oder ganzen Zahlen handeln.

In der englischsprachigen Literatur werden diese Merkmale häufig als *numerical* bezeichnet. Dieser Begriff ist aus Sicht des oben genannten Beispiels 2 jedoch irreführend, da kategoriale Merkmale durchaus numerischer Art sein können, ohne dass auf ihnen eine Metrik definiert ist.

Die Literatur kennt noch weitere Skalenniveaus für die genauere Unterscheidung quantitativer Merkmale. Diese sind für den weiteren Verlauf dieser Arbeit jedoch ohne Bedeutung und sollen daher nicht näher betrachtet werden.

2.1.3 *Häufigkeitsfunktionen*

Moderne Konzepte der Datenanonymisierung beziehen die Häufigkeiten von Merkmalen – ausgedrückt durch ihre Häufigkeitsfunktion – in ihre Funktionalität mit ein[3]. In der Statistik werden Häufigkeitsfunktionen dazu verwendet, um Informationen prägnanter darstellen oder zusammenfassen zu können.

Im englisch- sowie deutschsprachigen Raum wird die Häufigkeit des Auftretens einer Merkmalsausprägung x_i uneinheitlich mit h_i oder f_i bezeichnet [Mit12, S. 39]. Wir werden im späteren Verlauf der Arbeit vor das Problem gestellt, Häufigkeitsverteilungen von zwei Beobachtungen desselben Merkmals zu vergleichen. Um zwischen diesen unterscheiden zu können, ist es hilfreich, die Ausprägungen durch verschiedene Buchstaben darzustellen und sie über einen Index zu referenzieren. Diese Darstellung kann zu Konfusionen mit der oben genannten Notation der relativen und absoluten Häufigkeiten sowie der Merkmalsausprägungen führen, da sich diese visuell nicht stark genug voneinander unterscheiden. Wir folgen daher der Notation von Kohn wie sie nachfolgend beschrieben ist:

Definition 7 (Absolute Häufigkeit)**.** *Die Mächtigkeit der Menge der beobachteten Merkmalsausprägungen x_j, heißt absolute Häufigkeit und wird mit*

$$n(x_j) = \left|\{\omega_i | X(\omega_i) = x_j\}\right|, \; 1 \leq i \leq n, 1 \leq j \leq m \tag{2.2}$$

notiert.

3 Vgl. Kapitel 4.

Absolute Häufigkeiten hängen von der Anzahl der Beobachtungen n ab [Mit12]. Um uns zu ermöglichen Häufigkeitsverteilungen zueinander in Relation zu setzen, definieren wir nach Kohn den Terminus der *relativen Häufigkeiten*:

Definition 8 (Relative Häufigkeit)**.** *Die relative Häufigkeit beschreibt den Anteil der beobachteten Merkmalsausprägungen x_j am Gesamtumfang n der Stichprobe. Unter Beachtung der Voraussetzungen der absoluten Häufigkeit schreiben wir:*

$$f(x_j) = \frac{n(x_j)}{n} \quad . \tag{2.3}$$

Korollar 1. *Es folgt:*

$$0 \leq f(x_j) \leq 1 \quad . \tag{2.4}$$

Demnach gilt nach Kohn [Koh05]:

$$\sum_{j=1}^{m} f(x_j) = 1 \quad . \tag{2.5}$$

Die Folge der Zahlen $f(x_1), \ldots, f(x_m)$ wird *relative Häufigkeitsverteilung* genannt [Har09], sie lässt sich in Tabellen visualisieren. Im Kontext dieser Arbeit wird vornehmlich auf die relativen Häufigkeit des Auftretens einer Merkmalsausprägung referenziert werden. Wir führen daher die Konvention ein, dass eine *Häufigkeitsverteilung* stets eine Menge der relativen Häufigkeiten eines Merkmals darstellt. Wird eine Häufigkeitsverteilung im Fließtext beschrieben, so schreiben wir $X = \{f(x_1), \ldots, f(x_n)\}$, um darzustellen, dass es sich um die Häufigkeiten der Ausprägungen $x_1, \ldots, x_n$ des Merkmals X handelt. Werden zwei Häufigkeitsverteilungen desselben Merkmals miteinander verglichen, die aber aus unterschiedlichen Stichproben stammen, so versehen wir die Häufigkeitsverteilung mit einem Index, welcher die jeweilige Stichprobe kennzeichnet.

Beispiel 3. Es seien A und B zwei Stichproben des Merkmals X vom Umfang m und n entsprechend. Dann bezeichne

$$X_A = \{f_A(x_1), \ldots, f_A(x_m)\} \tag{2.6}$$

die Häufigkeitsverteilung des Merkmals X in der Stichprobe A und

$$X_B = \{f_B(x_1), \ldots, f_B(x_n)\} \tag{2.7}$$

die Häufigkeitsverteilung des Merkmals X in der Stichprobe B.

2.1.4 *Darstellung von Häufigkeitsverteilungen*

Die Häufigkeitsverteilung lässt sich unter anderem als *Säulendiagramm* darstellen. Dies ist die in dieser Arbeit bevorzugte Darstellungsform. Ein Beispiel findet sich auf Seite 46 in Abbildung 4.1c. Eine weitere Darstellungsform ist die des *Histogramms*. Diese Darstellungsform ist der des Säulendiagramms ähnlich. Abweichend werden jedoch die Daten der Übersichtlichkeit halber gruppiert. Laut Fahrmeir et al. geschieht diese Gruppierung nach dem *Prinzip der Flächentreue* [Fah+07, S. 42]. Dieses verlangt, dass die durch die gruppierten Säulen dargestellten Flächen proportional zu den absoluten oder relativen Häufigkeiten sind.

Dem der deskriptiven Statistik verwandten Gebiet der Wahrscheinlichkeitstheorie entstammt in diesem Zusammenhang der Begriff der *Wahrscheinlichkeitsmasse*. Als Wahrscheinlichkeitsmasse wird das Integral, d. h. die Fläche unterhalb der *Dichtekurve*, bezeichnet [Fah+07, S. 304]. Bei der Dichtekurve handelt es sich um eine stetige Funktion, mit deren Hilfe die Häufigkeitsverteilung approximiert wird [Fah+07, S. 87]. Wir werden diese Notation auf Säulendiagramme übertragen und in diesem Zusammenhang die relativen Häufigkeiten, welche durch die Säulen dargestellt werden, als Wahrscheinlichkeitsmasse bezeichnen.

2.1.5 *Mehrdimensionale Merkmale*

Neben der Betrachtung eines Merkmals einer Untersuchungseinheit, werden wir im Kontext dieser Arbeit zwei oder mehr Merkmale einer Untersuchungseinheit betrachten. Diese Merkmale werden zu einem Tupel zusammengefasst und als *mehrdimensionale* oder *multivariate* Merkmale bezeichnet [CK08, S. 14]. Um diese geeignet referenzieren zu können, werden wir ebenfalls die Darstellungsform eines Tupels für die Gesamtheit der Merkmale wählen. Wir schreiben $(X_1, X_2, \ldots, X_p)$ und bezeichnen mit p die *Dimension*, d. h. die Anzahl der Merkmale[4].

Das Ergebnis einer Stichprobe vom Umfang n ist ein *multivariater Datensatz* mit n Tupeln und p Dimensionen [CK08, S. 14].

Häufig werden die Ausprägungen aller Merkmale und Merkmalsträger tabellarisch dargestellt. Werden die Ausprägungen zweier Merkmale verglichen, so wird ebenfalls eine tabellarische Darstellung gewählt. Die entstehende Tabelle wird als *Häufigkeitstabelle* oder, je nach Skala des betrachteten Merkmals, als *Kontingenztafel* oder *Korrelationstabelle* bezeichnet [Koh05, S. 102]. Die tabellarische Darstellung er-

4 Der Übersichtlichkeit halber werden wir zumeist auf die Verwendung von verschiedenen Großbuchstaben zur Bezeichnung der Merkmale zurückgreifen. So ist (X, Y) ein zweidimensionaler Datensatz.

	MERKMAL X					
MERKMAL Y	x_1	$\dots$	x_i	$\dots$	x_k	SUMME
y_1	$n(x_1, y_1)$	$\cdots$	$n(x_i, y_1)$	$\cdots$	$n(x_k, y_1)$	$n(\cdot, y_1)$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_j	$n(x_1, y_j)$	$\cdots$	$n(x_i, y_j)$	$\cdots$	$n(x_k, y_j)$	$n(\cdot, y_j)$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_m	$n(x_1, y_m)$	$\cdots$	$n(x_i, y_m)$	$\cdots$	$n(x_k, y_m)$	$n(\cdot, y_m)$
SUMME	$n(x_1, \cdot)$	$\cdots$	$n(x_i, \cdot)$	$\cdots$	$n(x_k, \cdot)$	n

Tabelle 2.1: Kontingenztafel absoluter Häufigkeiten nach Kohn [Koho5]

möglicht die *gemeinsame Verteilung* zweier Merkmale hinsichtlich ihrer Abhängigkeit zu analysieren [Fah+07, S. 112].

Eine Kontingenztafel kann sowohl die absoluten Häufigkeiten des Auftretens der Kombination von Merkmalen beinhalten, als auch die relativen Häufigkeiten. Betrachten wir einen zweidimensionalen Datensatz (X, Y). Mit den Ausprägungen $\mathcal{A}_X = \{x_1, \dots, x_k\}$ sowie $\mathcal{A}_Y = \{y_1, \dots, y_m\}$. Es gilt: $i = 1, \dots, k, j = 1, \dots, m$ sowie $k, m \leq n$.

Die absolute Häufigkeit des gemeinsamen Auftretens zweier Merkmalsausprägungen $x_i \in X$ sowie $y_j \in Y$ wird durch $n(x_i, y_j)$ dargestellt. Entsprechend gilt für die relativen Häufigkeiten $f(x_i, y_j) = \frac{n(x_i, x_j)}{n}$.

Ein Beispiel für eine Kontingenztafel der absoluten Häufigkeiten des Auftretens der Merkmale X und Y ist in Tabelle 2.1 gegeben.

Eine Besonderheit der Kontingenztafel besteht im Inhalt der letzten Spalte sowie der letzten Zeile der Tabelle. Diese stellen die sogenannten *Randhäufigkeiten* dar und entsprechen dem absoluten (relativen) Auftreten der Spalten- bzw. Zeilenmerkmale. Dies wird durch die Punktnotation „$\cdot$“ der absoluten bzw. relativen Häufigkeitsfunktion angezeigt. Sie wird als Randsumme gelesen und entspricht $n(\cdot, y_j) = \sum_i n(x_i, y_j)$.

2.1.6 *Zusammenhangsmaße*

Der im vorangegangenen Abschnitt beschriebene Terminus der multivariaten Statistik hat vielfältige Analysemethoden empirischer Daten hervorgebracht. Für diese Arbeit von besonderer Bedeutung ist die des *Zusammenhangsmaßes*. Als Zusammenhangsmaß wird in der Statistik ein Verfahren bezeichnet, mit dessen Hilfe die

Beziehung zweier Merkmale untersucht werden kann. Eine Frage, welche im Verlauf dieser Ausarbeitung von Interesse sein wird, ist die der Unabhängigkeit zweier Merkmale. Dies wird einen immanenten Einfluss auf die Gestaltung eines verteilten Anonymisierungsalgorithmus haben, weshalb es an dieser Stelle notwendig ist, tiefer in das Themengebiet einzudringen. Wir definieren statistische Unabhängigkeit zweier Merkmale nach Kohn [Koho5] wie folgt:

Definition 9 (Statistische Unabhängigkeit zweier Merkmale)**.** *Zwei Merkmale X und Y mit den Ausprägungen $\mathcal{A}_X = \{x_1, \ldots, x_k\}$ sowie $\mathcal{A}_Y = \{y_1, \ldots, y_m\}$ heißen* statistisch unabhängig *genau dann, wenn gilt:*

$$f_{XY}(x_i, y_j) = f_X(x_i) \cdot f_Y(y_j) \tag{2.8}$$

Sind zwei Merkmale nicht unabhängig, stehen sie in einem Zusammenhang. Um die Stärke des Zusammenhangs zu messen, wurden Zusammenhangsmaße entwickelt. Für die Anwendung dieser muss unterschieden werden, welchem Messniveau ein beobachtetes Merkmal angehört. So kommen für nominale Merkmale der $\mathcal{X}^2$-Koeffizient oder der bekannte *Spearman Korrelationskoeffizient* für ordinal skalierte Merkmale in Betracht [Fah+07].

Das Problem dieser Zusammenhangsmaße ist einerseits ihre Spezialisierung. Zum anderen sind sie nur geeignet, einen *linearen* oder *monotonen* Zusammenhang zweier Merkmale zu quantifizieren [Koho5, S. 129]. Von einem linearen Zusammenhang wird gesprochen, sobald eine Erhöhung (Verringerung) einer Merkmalsausprägung zu einer linearen Erhöhung (oder Verringerung) der anderen Merkmalsausprägung führt. Dies bedeutet, dass der Grad der Erhöhung konstant ist. Bei einem monotonen Zusammenhang ist dies nicht der Fall, der Zusammenhang ist jedoch monoton steigend oder fallend [Fah+07, S. 137 & 143].

Abstandsmaße aus dem Bereich der Informationstheorie sind nicht von diesen Einschränkungen betroffen, weshalb wir diese genauer betrachten wollen.

2.2 VERWENDUNG DER INFORMATIONSTHEORIE

Der Begriff der *Information* sowie essentielle Grundlagen der modernen Informationstheorie basieren auf der 1948 verfassten Abhandlung „A mathematical theory of communication“ von Shannon [Wero8, S. 1].

Shannons Arbeit wurde durch die neuen Möglichkeiten der Signalübertragung und dem Fehlen von deren mathematischer Beschreibung motiviert.

Insbesondere erkannte er das Problem des Auftretens von Kommunikationsfehlern durch Rauschen und die Möglichkeit eines effizienten Datentransports durch

das Ausnutzen statistischer Eigenschaften des zu übertragenden Alphabets. [Sha48, S. 1]

Die Shannon'sche Informationstheorie ist für diese Arbeit von besonderer Relevanz. Im Verlauf der Arbeit werden wir feststellen, dass sich Daten eines zu anonymisierenden Datensatzes gegenseitig beeinflussen können. Eine Beurteilung dieses Einflusses mithilfe der statistischen Zusammenhangsmaße ist zuweilen umständlich, da die Ausprägungen der Daten unterschiedlicher Natur sein können. Ohne vorgreifen zu wollen, sei an dieser Stelle vermerkt, dass die Informationstheorie rege Verwendung bei der Entwicklung von Konzepten des Datenschutzes findet. Es lohnt sich daher an jener Stelle genauer auf sie einzugehen.

2.2.1 *Wahrscheinlichkeitsbegriff*

Bevor wir uns den zentralen Begriffen der Informationstheorie – *Entropie* und Information – widmen können, müssen wir uns zunächst mit dem Begriff der *Wahrscheinlichkeit* auseinandersetzen. Dieser bildet das theoretische Fundament des Häufigkeitsbegriffs, welchen wir in den vorherigen Abschnitten behandelt haben. Als Wahrscheinlichkeit wird die Häufigkeit des Eintretens eines ungewissen Ergebnisses im Rahmen eines *Zufallsexperiments* bezeichnet [Fah+07, S. 175]. Mithilfe der Wahrscheinlichkeit wird der Ausgang des Zufallsexperiments geschätzt. Dieses Zufallsexperiment kann das Ergebnis des dreifachen Werfens einer Münze darstellen oder jede andere beliebige Anwendung, deren Ausgang ungewiss ist.

In dem Zufallsexperiment bildet der *Ergebnisraum* die Menge der möglichen Ergebnisse eines Zufallsexperiments. Gewöhnlich wird der Ergebnisraum mit Ω bezeichnet. Die Ergebnisse werden entsprechend mit ω bezeichnet. Mehrere Ergebnisse werden als *Ereignisse* zusammengefasst. Beim vorher erwähnten Beispiel des dreimaligen Münzwurfs, wäre „Kopf, Kopf, Zahl" ein mögliches Ereignis. Ein Ereignis ist also eine Menge von Ergebnissen. Entsprechend nennt man einelementige Ergebnisse auch *Elementarereignisse*. Durch ihre Definition als Menge sind alle bekannten Mengenoperationen auf das Rechnen mit Wahrscheinlichkeiten übertragbar.

Bislang wurden die konkreten Realisierungen von Werten im Rahmen einer Stichprobe betrachtet. Aus diesen haben wir die absoluten und relativen Häufigkeiten gewonnen und Möglichkeiten genannt diese grafisch aufzubereiten. Wollen wir aufgrund der beobachteten Ereignisse Aussagen über zukünftige Ereignisse treffen, so müssen wir die Häufigkeiten in Wahrscheinlichkeiten überführen. Nach Fahrmeir et al. ergibt sich der sogenannte *objektivistische Wahrscheinlichkeitsbegriff* „aus einer Häufigkeitsinterpretation der Wahrscheinlichkeit" [Fah+07, S. 182]. Dies bedeutet, dass ab einer bestimmten Anzahl an Wiederholungen des Zufallsexperiments die

relativen Häufigkeiten der Ergebnisse den Wahrscheinlichkeiten entsprechen. Die Wahrscheinlichkeit des Eintretens eines Ereignisses A werden wir mit $P(A)$ bezeichnen.

Aus diesem, erstmals von Richard von Mises formulierten, Wahrscheinlichkeitsbegriff entwickelte Andrei Nikolajewitsch Kolmogorow die Axiome, auf denen die Wahrscheinlichkeitstheorie heute gründet [Haro9, S. 94]. Diese sind nach Fahrmeir et al. wie folgt definiert:

Definition 10 (Axiome von Kolmogorow)**.**

$$P(A) \geq 0 \tag{K1}$$
$$P(\Omega) = 1 \tag{K2}$$
$$A \cap B = \emptyset \implies P(A \cup B) = P(A) + P(B) \tag{K3}$$

Die Wahrscheinlichkeit P wird nach Fahrmeir et al. als eine Abbildung $P : \{A \subset \Omega\} \to [0, 1]$ aufgefasst [Fah+07, S. 182]. Aus Korollar 1 wissen wir bereits, dass die relative Häufigkeit des Auftretens eines Merkmals einen Wert zwischen 0 und 1 annimmt. Dies gilt entsprechend auch für die Wahrscheinlichkeit des Auftretens eines Ereignisses. Das *sicherere Ereignis* hat die Auftrittswahrscheinlichkeit 1, das *unmögliche Ereignis* die Wahrscheinlichkeit 0.

Sind zwei Ereignisse A und B paarweise disjunkt, so ist die Wahrscheinlichkeit für ihr *Vereinigungsereignis* als Summe ihrer Einzelwahrscheinlichkeiten anzusehen (K3). Für unser Münzwurfbeispiel bedeutet dies: Die Wahrscheinlichkeit nach einem Wurf das Ergebnis „Kopf“ oder „Zahl“ zu erreichen, ist gerade die Summe von $P(\text{Kopf})$ und $P(\text{Zahl})$. In diesem trivialen Fall ist das Ergebnis 1, denn es gilt: $P(\text{Kopf} \cup \text{Zahl}) = P(\text{Kopf}) + P(\text{Zahl}) = \frac{1}{2} + \frac{1}{2}$. Entsprechend ergibt sich das *Schnittereignis* $P(A \cap B)$ aus der Wahrscheinlichkeit des Auftretens von A und B.

Wir wollen abschließend das *bedingte Ereignis*, nach Fahrmeir et al. definieren [Fah+07, S. 202ff]:

Definition 11 (Bedingte Wahrscheinlichkeit)**.** *Die Wahrscheinlichkeit für A unter der Bedingung B ergibt sich als Quotient aus dem gemeinsamen Ereignis von A und B und der Wahrscheinlichkeit für das bedingende Ereignis.*

Seien $A, B \subset \Omega$ und $P(B) > 0$. Dann ist die bedingte Wahrscheinlichkeit *von A unter B definiert als*

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad .$$

2.2.2 *Information*

Wir beginnen die Einführung in die Informationstheorie mit der Klärung der Frage was *Information* bedeutet. Nach Shannon wird Information in Form von *Zeichen* aus einer Quelle zu einer Senke emittiert. Die Zeichen entstammen einem Alphabet an möglichen Zeichen und stellen somit die Elementarereignisse eines wahrscheinlichkeitstheoretischen Ergebnisraums dar. Die Bedeutung des Zeichens ist für die übertragene Information ohne Belang, vielmehr bestimmt die Wahrscheinlichkeit des Auftretens des Zeichens die von ihm transportierte Information [Wero8, S. 5]. So transportiert ein selten auftretendes Zeichen mehr Information als ein häufig auftretendes [Wero8, S. 4].

Shannon definierte wünschenswerte Eigenschaften einer Funktion zur Beschreibung der Information. Er wählte die Lochkarte als Beispiel eines Informationsträgers [Sha48, S. 1]. Verwendet man statt einer zwei Lochkarten, so quadriert dies die Anzahl der möglichen Nachrichten. Gleichzeitig sollte durch die Verdoppelung der Karten auch die doppelte Information transportiert werden können [Wero8, S. 1].

Die Lograrithmusfunktion ist eine Funktion, welche der beschriebenen Eigenschaft nachkommt und gleichzeitig mathematisch günstig ist [Sha48, S. 1]. So ist der Logarithmus einer Multiplikation zweier Faktoren äquivalent zur Addition der Logarithmen der Faktoren.

Ausgehend von den bisherigen Beobachtungen, lässt sich der Begriff des *Informationsgehalts* eines Zeichens – die durch das Zeichen transportierte Information – charakterisieren. Wir definieren nach Werner [Wero8, S. 4]:

Definition 12 (Informationsgehalt)**.** *Der Informationsgehalt I des Zeichens x_i eines Alphabets X mit der Wahrscheinlichkeit $P(x_i)$ beträgt*

$$I(P(x_i)) = -log_b(P(x_i)) \quad .$$

Üblicherweise beträgt die Basis b des Logarithmus 2. Ist dies der Fall, ordnen wir der Information die Kenngröße *bit* zu. Andere bekannte Kenngrößen sind, je nach Basis des Logarithmus, *decimal digits (dit)* oder *natural units (nat)* [Sha48, S. 1f].

Auf der Grundlage der vorangegangenen Definition der Information, ist es uns nun gestattet den Entropie-Begriff einzuführen.

2.2.3 *Entropie*

Nach Fey wird Information als „beseitigte Unsicherheit“ [Fey68, S.] verstanden.

Durch die Entropie wird der mittlere Informationsgehalt einer Nachrichtenquelle bezeichnet [Wero8, S. 5]. Auf eine genaue Definition der Nachrichtenquelle im Kon-

text der Informationstheorie sei an dieser Stelle verzichtet. Es genügt anzunehmen, dass die Quelle die Menge der Elementarereignisse sei.

Ausgehend von Werner definieren wir die Entropie wie folgt:

Definition 13 (Entropie)**.** *Sei X eine Nachrichtenquelle mit dem Zeichenvorrat $X = \{x_1, x_2, \ldots, x_n\}$ und den zugehörigen Wahrscheinlichkeiten $P(x_1), P(x_2), \ldots, p(x_n)$. Dann besitzt die Quelle die* Entropie *oder den* mittleren Informationsgehalt *von:*

$$\begin{aligned} H(X) &= \sum_{i=1}^{n} P(x_i) \cdot I\left(P(x_i)\right) = \sum_{i=1}^{n} P(x_i) \cdot -\log_b\left(P(x_i)\right) \\ &= \sum_{i=1}^{n} P(x_i) \cdot (-1) \cdot \log_b\left(P(x_i)\right) = -\sum_{i=1}^{n} P(x_i) \cdot log_b\left(P(x_i)\right) \quad . \end{aligned}$$

Nach Kohn liegt der Wertebereich der Entropie im Intervall $[0, log_b(n)]$ [Koho5, S. 42]. Das bedeutet, dass Ereignisse mit Wahrscheinlichkeit 0 keinen Beitrag zur Entropie liefern [Fey68]. Dem gegenüberstehend liefert die Nachrichtenquelle keine Information, sobald ein Zeichen x_i mit Wahrscheinlichkeit 1 auftritt. Denn dies bedeutet, dass dieses Zeichen mit absoluter Sicherheit auftritt und alle anderen nicht. Folglich wird keine Information übertragen, es ergibt sich die *minimale Entropie* von 0 [Fey68]. Die *maximale Entropie* wird erreicht, sobald jedes Ereignis gleich wahrscheinlich ist [Wero8, S. 5].

Die Informationstheorie ist eng mit der Wahrscheinlichkeitstheorie verwandt. Daher ergeben sich ähnlich komplexe Darstellungsmöglichkeiten für die Entropie. Wir werden im Folgenden jedoch ausschließlich die für den Kontext dieser Arbeit relevanten darstellen.

So ist die *Verbundentropie* nach Werner wie folgt definiert [Wero8, S. 34]:

Definition 14 (Verbundentropie)**.** *Gegeben zwei Nachrichtenquellen X und Y. Seien $x_i \in X : 1 \leq i \leq n$ sowie $y_j \in Y : 1 \leq i \leq m$ die von X und Y emittierten Zeichen. Dann ist die Verbundentropie definiert als:*

$$H(X, Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} P(x_i, y_j) \cdot \log_b\left(P(x_i, y_j)\right) \quad . \tag{2.9}$$

Sie beschreibt den Erwartungswert der Entropie der Zeichenpaare (x_i, y_j). Die Verbundentropie ist dienlich für die Berechnung der *bedingten Entropie*, d. h. die Information eines Zeichens x_i nach Auftreten eines Zeichens y_j. Wir wollen sie nach Werner [Wero8, S. 34] wie folgt ermitteln:

Definition 15 (Bedingte Entropie)**.** *Gegeben zwei Nachrichtenquellen X und Y. Seien $x_i \in X : 1 \leq i \leq n$ sowie $y_j \in Y : 1 \leq i \leq m$ die von X und Y emittierten Zeichen. Dann ist die bedingte Entropie definiert als:*

$$H(X|Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j) \cdot \log_b \left(P(x_i|y_j)\right) \quad . \tag{2.10}$$

Beide Größen – Verbundentropie sowie bedingte Entropie – ermöglichen die Ermittlung eines Zusammenhangsmaßes. Gemeint ist die sogenannte *Transinformation*, welche auch als *Wechselseitiger Informationsgehalt* bekannt ist. Sie beschreibt die Korrelation zweier Zufallsgrößen [Wero8, S. 90f].

Definition 16 (Transinformation)**.** *Gegeben zwei Nachrichtenquellen X und Y. Seien $x_i \in X : 1 \leq i \leq n$ sowie $y_j \in Y : 1 \leq i \leq m$ die von X und Y emittierten Zeichen. Dann ist die Transinformation definiert als:*

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j) \cdot \log_b \left(\frac{P(x_i, y_j)}{P(x_i)P(y_j)}\right) \quad . \end{aligned}$$

Nach Kohn liegt der Wertebereich der Transinformation im Intervall $[0, \min(H(X), H(Y))]$ [Koho5, S. 118]. Um einen normierten Wert im Intervall $[0, 1]$ zu erhalten, definieren wir anschließend die *normierte Transinformation* nach Kohn [Koho5, S. 118]:

Definition 17 (Normierte Transinformation)**.** *Sei die Transinformation zweier Nachrichtenquellen $I(X;Y)$ gegeben. Dann berechnet sich die* normierte Transinformation *$NI(X;Y)$ durch:*

$$NI(X;Y) = \frac{I(X;Y)}{\min(H(X), H(Y))} \quad . \tag{2.11}$$

Nach obiger Definition sprechen wir im Fall $NI(X;Y) = 1$ von einer *perfekten statistischen Abhängigkeit* der beiden Merkmale X und Y [Koho5, S. 119].

Im nachfolgenden Kapitel wollen wir den Terminus des „statistischen Merkmals“ in den Kontext der Datenanonymisierung überführen.

3 GRUNDLAGEN DER ANONYMISIERUNGSKONZEPTE

Im vorherigen Kapitel haben wir uns umfassend mit der dieser Arbeit zugrunde liegenden Mathematik beschäftigt. Dabei haben wir eine Terminologie eingeführt, welche für den Kontext der beschreibenden Statistik geschaffen wurde.

Bevor wir uns den Konzepten der Datenanonymisierung widmen, benötigen wir eine entsprechende Terminologie. Diese wollen wir nachfolgend einführen und mit den Begriffen des vorherigen Kapitels verknüpfen. Anschließend werden wir Datenstrukturen besprechen, welche für die Anonymisierungskonzepte wesentlich sind.

Die Übersetzung der englischsprachigen Begriffe aus dem Bereich der Relationalen Datenbanken ist, soweit nicht anders angegeben, Saake et al. [SHoo] entnommen. Begriffe, die aus dem Bereich der Anonymisierung von Daten stammen, wurden nach Karjoth [Karo8] ins Deutsche übersetzt.

3.1 DATENREPRÄSENTATION

Diese Diplomarbeit beschäftigt sich mit der Anonymisierung von Daten. Um die Mechanismen zum Erreichen dieses Ziels zu verstehen, müssen wir zunächst definieren was *Daten* im Sinne dieser Diplomarbeit bedeuten. Im Kontext dieser Arbeit wird vornehmlich über die Verarbeitung sogenannter *Mikrodaten* gesprochen. Mikrodaten beschreiben dabei nicht die Repräsentation der Daten als Relation, Tabelle oder Objekt, sondern lediglich die Tatsache, dass es sich bei den Daten um sämtliche Informationen einer *Datenerhebung* handelt [De +11, S. 2]. Eine Datenerhebung beschreibt beispielsweise eine Stichprobe (vgl. Kapitel 2). Die Daten sind nicht aggregiert, weshalb sie für gewöhnlich einen hohen Detaillierungsgrad besitzen [Mac+07; Karo8]. Mikrodaten sind in der Regel *personenbezogen*, d. h. sie entstammen einer allgemein bekannten und zugänglichen Population, bestehend aus mehreren *Individuen*. Die Population wird, im Einklang mit Abschnitt 2.1.1, mit Ω bezeichnet. Ein Individuum der Population Ω entspricht demnach einem Merkmalsträger. Wir wollen es in Übereinstimmung mit Kapitel 2 mit ω bezeichnen. Die Mikrodaten stellen demnach eine Stichprobe $\mathcal{X}$ der Population Ω dar.

Wir definieren nach Karjoth [Karo8]:

Definition 18 (Mikrodaten)**.** *Mikrodaten bezeichnen Daten über Individuen, welche, im Rahmen einer statistischen Erhebung, aus einer Stichprobe $\mathcal{X}$ der Grundgesamtheit Ω er-*

fasst wurden. Die Daten liegen in unveränderter, nicht-aggregierter Form vor. Weiterhin enthalten sie keine Informationen über die Struktur der durch sie abgebildeten Daten.

Für den Rest dieser Diplomarbeit legen wir fest, dass jedes Individuum $\omega_i, i \in \mathbb{N}$ genau einmal in der Stichprobe vorhanden ist. Dies ist eine häufige Annahme, wie sie unter anderem von Sweeney [Swe02b, S. 6] getroffen wird.

Typischer Weise werden Mikrodaten *tabellarisch* dargestellt und auch die in dieser Arbeit vorkommenden Beispiele werden sich dieser Darstellungsform bedienen. Eine *Tabelle* beschreibt die Repräsentation von Mikrodaten in Zeilen und Spalten. Sie enthält Daten über Individuen, welche zeilenweise als Einträge organisiert sind. Jede Zeile wird durch ein *Tupel* beschrieben. Dieses umfasst eine Menge von *Attributen* $\{A_1, \ldots, A_n\}, n \in \mathbb{N}$, welche mit dem Individuum assoziiert sind.

Jedes Attribut $A_i \in \{A_1, \ldots, A_n\}$ wird durch eine Spalte der Tabelle beschrieben. Die Spalte bildet als Menge von Werten aus einer *Domäne* eine semantische Einheit. Bezogen auf die Notation des vorhergegangenen Kapitels ist der Begriff des Attributs äquivalent zu dem des Merkmals. Der Begriff der Domäne lässt sich auf den des Wertebereichs aus dem vorherigen Kapitel abbilden.

Der Terminus der Tabelle entspricht dem Begriff der *Relation* des Relationenmodells relationaler Datenbank-Management-Systeme (DBMS). [Swe02b, S. 6]. Wir wollen Mikrodaten im Folgenden im Kontext relationaler Datenbank-Management-Systeme betrachten. Die Repräsentation der Mikrodaten als Relation ist jedoch keinesfalls zwingend [Sam01, S. 1012].

Um über eine einheitliche Notation zu verfügen, definieren wir nachfolgend den Begriff des Tupels nach Ullman [Ull88, S. 43]:

Definition 19 (Tupel)**.** *Sei $\{A_1, \ldots, A_n\}$ die Menge der durch die Mikrodaten repräsentierten Attribute. Bezeichne ferner $dom(A_i)$ die Domäne des Attributs $A_i \in \{A_1, \ldots, A_n\}$. Ferner sei die* Relation *R eine Teilmenge des Kreuzprodukts der Domänen $\{dom(A_1) \times \ldots \times dom(A_n)\}$.*

*Dann heißt ein Element $t \in R$ n-*Tupel *oder kurz* Tupel*. Das Tupel wird durch die Ausprägungen seiner Attributwerte, die Instanz von R, vollständig beschrieben.*

Ferner beschreibe $t[A_i, \ldots, A_j]$ die Werte des Tupels t in den Attributen $A_i, \ldots, A_j$.

Im Kontext dieser Arbeit werden die Begriffe *Individuum*, *Eintrag* und *Tupel* synonym verwendet. Wir definieren eine Tabelle nach Sweeney [Swe02b, S. 6] wie folgt:

Definition 20 (Tabelle)**.** *Eine Tabelle $T(A_1, \ldots, A_n)$ mit den Attributen A_1 bis $A_n, n \in \mathbb{N}$ ist definiert über ihr Schema R sowie die Menge ihrer momentanen Instanzen, d. h. die Menge der Tupel, welche dem Schema genügen.*

Die Anzahl der Tupel einer Tabelle sei mit $|T|$ beschrieben.

Eine Tabelle stellt die Grundlage einer Veröffentlichung dar. Wir werde die Tabelle auch als Veröffentlichung bezeichnen. Bereits in Kapitel 1 wurde die Notwendigkeit einer Anonymisierung der Tabelle vor ihrer Veröffentlichung erläutert. In diesem Kontext wurde auf das Mittel der De-Identifikation zum Schutz der Daten verwiesen. Anhand der bekannten Studie von Sweeney wurde auf die Gefahr der Re-Identifikation durch Linking hingewiesen [Sweoo]. Die durch das Linking entstehende Bedrohung wird als *Datenverknüpfungsproblem* bezeichnet [Karo8, S. 19]. Es wurde auf bestimmte Attribute verwiesen durch deren Existenz sich das Datenverknüpfungsproblem ergibt, ohne diese genauer zu benennen. Nachfolgend wollen wir die Attribute einer Tabelle klassifizieren, um uns dem Datenverknüpfungsproblem formal zu nähern. Zu diesem Zweck definieren wir die *Projektion* einer Tabelle nach Ullman [Ull88, S. 56] wie folgt:

Definition 21 (Projektion). *Sei $T(A_1, \ldots, A_n)$ eine Tabelle mit Attributen $\{A_1, \ldots, A_n\}$. Dann bezeichnet $\Pi_{A_i,\ldots,A_j}(T) = \{t[A_i, \ldots, A_j] : t \in T\}$ die Projektion von T auf die Attribute A_i bis A_j, mit $\{A_i, \ldots, A_j\} \subseteq \{A_1, \ldots, A_n\}$.*

Es entspricht unserer Intuition, dass einige Attribute schützenswerter erscheinen als andere. Auf dieser Grundlage teilen wir die Attribute $\{A_i : 1 \leq i \leq n\}$ in vier Klassen ein:

1. Identifikator
2. Quasi-Identifikator
3. Sensibles Attribut
4. Neutrales Attribut

Wir wollen zunächst versuchen eine Intuition für die Klassifizierung der Attribute zu geben.

Attribut-Typ 1 wird als *Identifikator (ID)* bezeichnet. Nach Karjoth ist ein ID wie folgt definiert: „Ein Identifikator ist ein Attribut, welches eine Person eindeutig identifiziert." [Karo8, S. 20] Ein Beispiel hierfür ist die Sozialversicherungsnummer. Aber auch der Name kann als ID dienen. Daher wird er für gewöhnlich aus einer Veröffentlichung entfernt [Samo1, S. 1012]. Dieser Vorgang ist uns bereits aus Kapitel 1 als De-Identifikation bekannt.

Dem als *Quasi-Identifikator (QID)* bezeichneten Attribut-Typ in Klasse 2 kommt eine Sonderstellung zu. Nach Dalenius bezeichnen wir Attribute, welche für das Datenverknüpfungsproblem anfällig sind, als QID [Dal86]. Die Klassifikation eines Attributs als QID ergibt sich aus der Existenz dieses Attributs in weiteren veröffentlichten Tabellen. Ein klassisches Beispiel, welches auch Sweeney für ihre Studie wählte, sind

Identifikator	Quasi-Identifikator	neutrales Attribut	Sensibles Attribut	
NAME	GESCHLECHT	NEUTR. ATTRIBUT	DIAGNOSE	} Attribute
Name_1	männlich	Attributwert_1	Diagnose_1	} Tupel / Individuum / Eintrag
Name_2	weiblich	Attributwert_2	Diagnose_2	
...	...	...	...	
Name_n	weiblich	Attributwert_n	Diagnose_n	

$T(\text{NAME,GESCHLECHT,NEUTR. ATTRIBUT,DIAGNOSE})$

Tabelle 3.1: Veranschaulichung der tabellarischen Repräsentation von Mikrodaten

die demographischen Attribute „Alter“, „Geschlecht“ und „Postleitzahl“ [Swe00, S. 2]. Mit ihrer Hilfe konnte Sweeney im Jahre 2000 87 % der amerikanischen Bevölkerung durch den Abgleich von Wählerlisten und Krankenhausdaten re-identifizieren [Swe00, S. 2]. Eine weitere Studie kam zu dem Ergebnis, dass anhand dieser Attribute 61 % der Bevölkerung der USA re-identifiziert werden konnte [Gol06]. Dieser Umstand verdeutlicht das Problem der Definition des QID, da er von der Existenz weiterer veröffentlichter Tabellen abhängt. Sweeney konnte jedoch zeigen, dass mit steigender Anzahl von QIDs die Wahrscheinlichkeit der Re-Identifikation steigt [Swe00, S. 2].

Attribut-Typ 3 – das *Sensibles Attribut (SA)* – repräsentiert „Attribute, welche Informationen über eine Person darstellen, die nicht mit ihr verbunden werden sollen“ [Kar08, S. 20]. Bei diesen kann es sich um das Gehalt oder den Gesundheitszustand eines Individuums handeln. Diese Klasse der Attribute stellt denjenigen Typ dar, dessen Bekanntwerden durch Linking zu einer Verletzung der Privatsphäre führen würde.

Attribut-Typ 4 stellt einen Sammelbegriff für sämtliche Attribute dar, welche nicht durch eine der vorherigen Klassen repräsentiert werden. Wenngleich einige Arbeiten diesen Typ betrachten [BS08], so spielt er doch für den weiteren Verlauf dieser Arbeit keine Rolle und sei nur der Vollständigkeit halber erwähnt.

Tabelle 3.1 visualisiert die uns bekannte Notation sowie die vorgestellten Klassen.

Anhand dieser werden wir die vorher genannten Begriffe des ID, QID und SA anhand der in Tabelle 3.2 genannten Daten konkretisieren. Im Folgenden werden die Attribute der Tabelle 3.2 im Fließtext in Schreibmaschinenschrift gesetzt.

Wir definieren den Begriff des ID wie folgt:

ID	NAME	PLZ	ALTER	GESCHL.	GEHALT (*in €*)	KRANKHEIT	L-WBC (Abs. Zahl/*ml*)	CRP (*Referenzbereich*)
1	Anton	47677	29	46,XY	3000	Magengeschw.	11000	erhöht
2	Bert	47602	22	46,XY	4000	Gastritis	10000	stark erhöht
3	Cäsar	47678	27	46,XY	5000	Magenkrebs	9000	erhöht
4	Dora	47905	43	45,Xo	6000	Gastritis	8000	erhöht
5	Emilie	47909	49	46,XX	11000	Grippe	3000	normal
6	Friederike	47906	47	46,XX	8000	Bronchitis	6000	leicht erhöht
7	Gustav	47605	30	47,XXY	7000	Bronchitis	7000	leicht erhöht
8	Heinrich	47673	36	46,XY	9000	Magenkrebs	5000	erhöht
9	Ingo	47607	32	46,XY	10000	Lungenentz.	4000	normal

Tabelle 3.2: Verwendete Daten

Definition 22 (Identifikator). *Sei eine Tabelle $T(A_1, \ldots, A_n)$ mit den Attributen $\{A_1, \ldots, A_n\}$ gegeben. Ferner bezeichne ID_T die Menge der Identifikatoren aus T, mit $ID_T \subseteq \{A_1, \ldots, A_n\}$. Dann gilt für den Kontext dieser Arbeit:*

$$ID_T := \{\texttt{ID}, \texttt{Name}\} \tag{3.1}$$

Der Identifikator bezeichnet eine eindeutige Nummer sowie den Namen eines Individuums. In den begleitenden Beispielen werden wir der Einfachheit halber ausschließlich das Attribut `ID` anführen. Da dieses durch die De-Identifikation nie Teil einer veröffentlichten Tabelle sein wird, ist es in den Beispielen grau dargestellt. Es dient lediglich der Orientierung und dem besseren Verständnis der Beispiele.

Wir wollen zunächst den Begriff des SA definieren, bevor wir uns dem QID zuwenden.

Definition 23 (Sensibles Attribut). *Sei eine Tabelle $T(A_1, \ldots, A_n)$ mit den Attributen $\{A_1, \ldots, A_n\}$ gegeben. Ferner bezeichne SA_T die Menge der sensiblen Attribute aus T, mit $SA_T \subseteq \{A_1, \ldots, A_n\}$. Dann gilt für den Kontext dieser Arbeit:*

$$SA_T := \{\texttt{Gehalt}, \texttt{Krankheit}, \texttt{L-WBC}, \texttt{CRP}\} \tag{3.2}$$

Die Attribute `Leukozyten der Gehirn-Rückenmarks-Flüssigkeit (L-WBC)` und `C-reaktives Protein (CRP)` sind Messwerte, welche in medizinischen Studien erhoben werden. Sie sind Indikatoren für Aktivitäten des Immunsystems, welche auf eine Infektion hindeuten [Hero8]. Sie werden in den uns begleitenden Beispielen Verwendung finden. Zusammen mit den Attributen `Gehalt` und `Krankheit` wurden sie derart gewählt, dass sämtliche Skalenniveaus in den SAs der Tabelle vertreten sind (vgl. Kapitel 2).

Wir haben bereits gesehen, dass sich die Definition des QID aus weiteren veröffentlichten Tabellen ergibt. Zur Formalisierung des QID definieren wir daher zunächst zwei Hilfsfunktionen:

$$f_c : \mathcal{X} \to T \tag{3.3}$$
$$f_g : T \to \Omega \tag{3.4}$$

Die Intuition hinter diesen Abbildungen sei wie folgt: Die Funktion f_c ermöglicht uns die Darstellung der Mikrodaten als Tabelle, d. h. die Zuordnung eines Merkmalsträgers $\omega \in \mathcal{X}$ zu einem Tupel $t \in T$. Demgemäß überführt Funktion f_c Einträge von dem Wertebereich der Stichprobe in den Wertebereich der Tabelle – das Schema. Die Funktion f_g beschreibt eine Abbildung, welche Tupel der Tabelle $t \in T$ auf Individuen der Population $\omega \in \Omega$ abbildet. Intuitiv stellt f_g die Zuordnung von Einträgen der Tabelle zu realen Personen dar.

Nach Sweeney [Swe02b, S. 7] ist der QIDs folgendermaßen definiert:

Definition 24 (Quasi-Identifikator)**.** *Gegeben eine Stichprobe $\mathcal{X}$ sowie eine Tabelle $T(A_1, \ldots, A_n)$. Ein Quasi-Identifikator von T, geschrieben als QI_T, ist eine Menge von Attributen $\{A_i, \ldots, A_j\} \subseteq \{A_1, \ldots, A_n\}$ aus der Tabelle T mit der Eigenschaft, dass*

$$\exists \omega_i \in \mathcal{X} : f_g\left(\Pi_{QI_T}(f_c(\omega_i))\right) = \omega_i \quad .$$

Demnach beschreibt ein QID eine Menge von Attributen mit der Eigenschaft, dass ein Tupel in der Tabelle existiert, welches bezüglich seiner QIDs ein Individuum der Population eindeutig kennzeichnet.

Im Rahmen dieser Arbeit verfügen wir über keine derartige Population. Wir definieren daher die Menge QI_T bezogen auf Tabelle 3.2 wie folgt:

$$QI_T := \{\texttt{PLZ}, \texttt{Alter}, \texttt{Geschlecht}\} \tag{3.5}$$

Es sei an dieser Stelle auf die Besonderheit des Attributs `Geschlecht` hingewiesen. Es wird in den nachfolgenden Beispielen durch seinen *Karyotyp*, d. h. die Anzahl der Chromosomen eines Individuums inklusive der Ausprägung der Geschlechtschromosomen, dargestellt. Im Kontext dieser Arbeit bezeichnet es kein sensibles Attribut und wurde aus Gründen der Variabilität der Anonymisierung gewählt. Ein Karyotyp von „46,XX“ bezeichnet eine gesunde Frau. Ein Karyotyp von „46,XY“ kennzeichnet einen gesunden Mann. Die Ausprägungen „45,Xo“ stellt den Karyotyp einer Frau dar, welche unter dem Turner-Syndrom leidet. Der Karyotyp von „47,XXY“ entspricht einem Mann, bei dem das Klinefelder-Syndrom diagnostiziert wurde. Die letztgenannten Ausprägungen stellen geschlechtsspezifische Erkrankungen dar.

Abschließend sei darauf hingewiesen, dass die Verknüpfung von Tabellen über ihre QIDs semantisch der Join-Operation der relationalen Algebra entspricht. Der Begriff lässt sich jedoch nicht direkt übertragen, da die Datenstrukturen der Mikrodaten und insbesondere der Population Ω unbestimmt sind [Sam01]. Wir werden im Folgenden weiterhin von der Verknüpfung von Tabellen mittels Linking sprechen.

Es bleibt anzumerken, dass in realen Szenarien bereits die Möglichkeit ein Individuum nahezu eindeutig einem Eintrag der Tabelle zuordnen zu können, bereits eine Gefahr für die Privatsphäre darstellen kann.

Wir haben die Grundlagen zum Verständnis der Anonymisierungskonzepte gelegt. Nachfolgend wollen wir die Prinzipien erläutern auf denen die in dieser Arbeit besprochenen Konzepte beruhen.

3.2 PRINZIP DER GENERALISIERUNG UND UNTERDRÜCKUNG

Bislang haben wir gesehen, dass die Attribute einer Tabelle unterschiedliche Rollen nach Veröffentlichung einer Tabelle annehmen. Als einzige Maßnahme zum Schutz vor Linking ist uns das Mittel der De-Anonymisierung bekannt.

Die Anonymisierung durch Entfernen explizit identifizierender Attribute ist nicht die einzige Möglichkeit des Schutzes einer Tabelle. Es existiert eine Vielzahl an Schutzmechanismen. So beschäftigen sich Arbeiten mit der Entwicklung von DBMS nach dem Prinzip *Privacy by Design (PbD)*. Agrawal et al. [Agr+02] skizzierten einen Entwurf für ein Privatsphäre-achtendes DBMS: Die *Hippocratic Database*.

Die Mehrzahl der Mechanismen zum Schutz sensibler Daten findet ihren Ursprung in Arbeiten zu statistischen Veröffentlichungen [Fun+10, S. 14:6]. Diese Konzepte vereinen sich unter dem Begriff des *Privacy-Preserving Data Publishing (PPDP)*. Der Schutz von Mikrodaten vor deren Veröffentlichung ist ein relativ junges Forschungsgebiet, welches sich ebenfalls dem Begriff des PPDP unterordnet [Fun+10, S. 14:6]

Nach Ciriani et al. können die Prinzipien des PPDP zum Schutz von Mikrodaten in zwei Kategorien eingeteilt werden [Cir+07, S. 8f]:

1. Erstellung synthetischer Daten
2. Datenmaskierung

Die Verwendung des Prinzips der Erstellung synthetischer Daten, beschreibt die Ersetzung der Tabelle mit einer künstlich erzeugten Tabelle. Diese erhält die statistischen Eigenschaften der Originaldaten [Cir+07, S. 9]. Wir werden diese Kategorie im Folgenden nicht weiter betrachten. Sie sei der Vollständigkeit halber erwähnt.

Das Prinzip der Datenmaskierung beinhaltet Techniken, welche auf den Originaldaten der Attributwerte einer Tabelle operieren. Zu diesen zählt das Verändern der Attributwerte (*Data Pertubation*) durch Hinzufügen von *statistischem Rauschen (Random Noise)* [Cir+07, S. 15f]. Statistisches Rauschen beschreibt die Veränderung der sensiblen Attributwerte durch Addition oder Multiplikation eines zufälligen Wertes. Diese Veränderung darf die statistischen Eigenschaften der Verteilung der Attributwerte nicht beeinflussen. Eine weitere Möglichkeit ist das, als *Swapping* bezeichnete, Vertauschen von Attributwerten zwischen den Tupeln. Die Methoden der Data Pertubation verfälschen die Originaldaten [Sam01, S. 1011]. Medizinische Studien erfordern jedoch einen Erhalt der Semantik der Daten. Aufgrund dieser Tatsache und des relativ geringen Schutzes der durch die Methoden der Data Perturbation erlangt wird, eignen sie sich nicht für eine weitere Betrachtung.

Einen weiteren Mechanismus zum Schutz der Daten durch Maskierung stellt das Prinzip der *Generalisierung und Unterdrückung* dar. Unter Generalisierung wird die Veränderung eines quasi-identifizierenden Attributwerts in einen weniger spezifischen, aber semantisch geeigneten Wert verstanden. Die Unterdrückung eines Attributwerts beschreibt das Entfernen oder Ersetzen des Wertes durch ein Zeichen, welchem im Kontext der Domäne des Attributs keine Bedeutung zukommt [Swe02a, S. 2]. Die Unterdrückung eines Attributs, d. h. all seiner Instanzen, entspricht der De-Anonymisierung [MBM11].

Die Gefahr von Linking, durch die Existenz von QIDs, wurde durch den Statistiker Dalenius bereits im Jahre 1986 erkannt. Zum Schutz gegen Linking wurden Methoden der Generalisierung und Unterdrückung entwickelt. Diesen fehlte jedoch eine einheitliche Grundlage. Da sich die Vorgehensweise der Generalisierung als wirkungsvoll erwies, formalisierte Sweeney das Prinzip des PPDP durch Generalisierung und Unterdrückung [Swe02a, S. 2].

In diesem Kontext sei auf den Begriff des *Privacy-Preserving Data Mining (PPDM)* hingewiesen. Dieses Forschungsfeld integriert die Konzepte des PPDP in Algorithmen des Data Mining (DM). In neueren Fassungen des Begriffs umschließt dieser ebenfalls Techniken zum Schutz von statistischen Datenbanken (vgl. [KE11, S. 379]) durch Konzepte der Anfrage-Beschränkung [Fun+10, S. 14:5f]. Wir werden uns im Rahmen dieser Diplomarbeit nicht mit den Konzepten des PPDM befassen. Sämtliche Konzepte des Datenschutzes, welche im Kontext dieser Arbeit besprochen werden, bauen auf dem Prinzip des PPDP durch Generalisierung und Unterdrückung auf. Zu diesem Zweck werden wir uns zunächst mit den Datenstrukturen befassen, auf denen dieses Prinzip beruht.

3.2.1 *Domänen-Hierarchie*

Zu Beginn dieses Kapitels wurde der Begriff des Tupels über die Instanz, d. h. die Ausprägung seiner Attributwerte, definiert. Der Begriff der Domäne ist an dieser Stelle wesentlich. Ein Relationsschema des Relationenmodells relationaler DBMS ist über den Relationsnamen und einer Menge von 2-Tupeln bestehend aus dem Attributsnamen sowie der zugehörigen Domäne definiert [Ull88]. Diese Definition entspricht der Definition der Tabelle aus Definition 20.

Die Domäne beschreibt den Wertebereich des Attributs. Dieser könnte im Falle der Postleitzahlen eine fünfstellige Kombination aus natürlichen Zahlen kleiner 10 darstellen. Ein Beispiel stellen die Postleitzahlen „47602“ und „47605“ aus 3.2 dar. Dem Prinzip der Generalisierung folgend, wäre eine semantisch korrekte, jedoch weniger spezifische Postleitzahl von „4760∗“ denkbar. Dies würde den vergrößerten Postleitzahlenbereich von „47600“ bis „47609“ darstellen.

Durch den starren Begriff der Domäne ist es uns jedoch nicht möglich diese Generalisierung in der Tabelle abzubilden [Swe02a, S. 5]. Daher erweiterte Sweeney den Begriff der Domäne um eine Domänen-Hierarchie [Swe02a, S. 5].

Eine Domänen-Hierarchie beschreibt die Übergänge von einer spezifischen Domäne zu einer weniger spezifischen, aber semantisch konsistenten Domäne. Die Domänen-Hierarchie wird als *gewurzelter Baum* dargestellt [Die10, S. 15f]. Der Baum sei von den Blättern hin zur Wurzel gerichtet. Die Domänen sind durch die Knoten des Baums repräsentiert. Die *Generalisierung* eines Attributwerts wird durch den Übergang – die Kante – eines Knoten zu einem anderen Knoten dargestellt. Syntaktisch beschreibt die Generalisierung die Ersetzung des Attributwerts aus einer Ebene des Baums durch einen Wert der nächst-höheren Ebene. Die Ebene der Originaldaten wird als Ground Domain H_0 bezeichnet [Swe02a, S. 5]. Sie findet sich in den Blättern des Baums. Die Ebene der msind aximalen Generalisierung wird als H_{max} gekennzeichnet. Sie bildet die Wurzel des Baums. Der Index $i \in \mathbb{N}$ beschreibt die Höhe der Ebene H_i ausgehend von den Blättern.

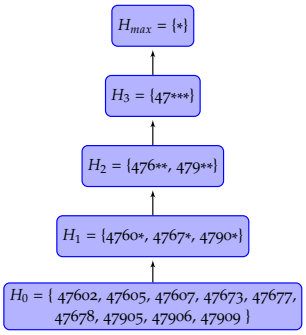
Eine Datenstruktur, welche eine Domänen-Hierarchie abbildet, ist die *Domain-Generalization-Hierarchy (DGH)* aus Abbildung 3.1a. Sie Beschreibt die Domänen-Hierarchie des Attributs `PLZ` aus Tabelle 3.2.

In der Abbildung 3.1a kennzeichnet bereits die Ebene H_3 eine maximale Generalisierung, da in ihr nur noch ein Wert möglich ist: Die Postleitzahl „47∗∗∗“. Die Ebene H_{max} beschreibt die *Unterdrückung* des Attributwerts, durch die Ersetzung der Domäne mit dem Zeichen „∗“ . Auf die Darstellung der DGHs für die Attribute `Alter` und `Geschlecht` sei an dieser Stelle verzichtet.

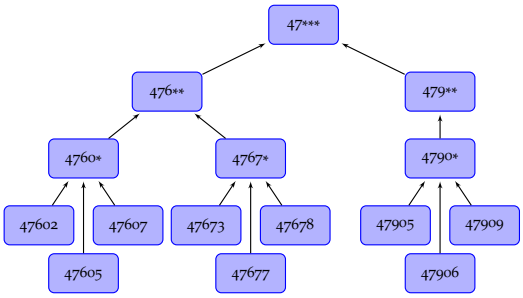
Durch die Bildung eines DGH ist es uns möglich die Domäne der jeweiligen Generalisierung anzupassen. Bislang fehlt uns jedoch eine Vorschrift, wie wir einen Wert

der Ebene H_i in einen Wert der Ebene H_{i+1} überführen. Dieses Ziel erreichen wir mithilfe einer verwandten Datenstruktur: Der *Value-Generalization-Hierarchy* (VGH) [Swe02b, S. 5]. In diesem Baum wird jedem Wert einer Hierarchie-Ebene ein separater Knoten zugeordnet. Die Ebene des Knotens bleibt im Vergleich zum DGH unverändert. Die Werte der Ground Domain bilden die Blätter des Baums. Die maximale Generalisierung oder Unterdrückung bildet die Wurzel.

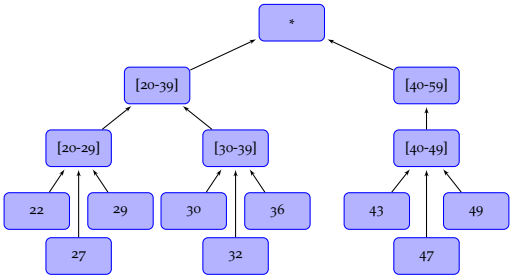
Ein VGH kann unterschiedlich konstruiert werden. Speziell für quantitative Attributwertsdomänen eignet sich die Verwendung einer *Partitionierung der Domäne* [LDR06, S. 2]. Diese teilt die Domäne in nicht überlappende Intervalle. Eine derartige Partitionierung kann ab Ebene 1 des VGH für das Attribut `Alter` (vgl. Abbildung 3.1c) betrachtet werden. Eine weitere Möglichkeit besteht in der Verwendung statischer Hierarchien. Diese Vorgehensweise ist speziell für kategoriale Attributwertsdomänen geeignet. Beispiele hierfür finden sich in den VGHs der Attribute `Geschlecht` sowie `PLZ` (vgl. Abbildungen 3.1d sowie 3.1b).



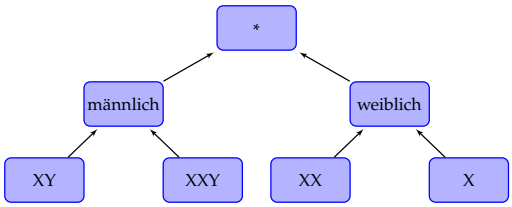
(a) DGH des Attributs `PLZ`



(b) VGH des Attributs `PLZ`



(c) VGH des Attributs `Alter`



(d) VGH des Attributs `Geschlecht`

Wir haben uns Werkzeuge geschaffen mit deren Hilfe wir den Begriff der Domäne um eine Hierarchie erweitert haben. Durch diese wird eine Generalisierung oder Unterdrückung von quasi-identifizierenden Attributwerten ermöglicht.

4 ZENTRALISIERTE ANONYMISIERUNGSKONZEPTE

> „The goal of privacy preservation is not only to protect privacy, but also to make the published data useful as much as the microdata."
>
> — Liu et al. [LLH10, S. 2]

Dieses Zitat beschreibt die Intuition hinter den Konzepten des PPDP. Durch das Prinzip der Generalisierung bleibt die Semantik der Daten erhalten. Dies ist bei anderen Konzepten des PPDP, wie dem der Data Perturbation nicht gegeben (vgl. Kapitel 3).

Im vorherigen Kapitel konnten wir die Prinzipien betrachten, welche den Konzepten des PPDP zu Grunde liegen. In diesem Kapitel wollen wir uns mit dem bekanntesten Konzept dieser Gattung auseinandersetzen: Der k-Anonymisierung. Die k-Anonymisierung bildet die Grundlage für fortgeschrittene Konzepte des PPDP, welchen wir uns im Laufe dieses Kapitels widmen werden.

Das Konzept der k-Anonymisierung ist von herausragender Wichtigkeit, da mit ihm erstmalig eine Formalisierung des Prinzips der Generalisierung und Unterdrückung erfolgte [Sam01, S. 1011].

4.1 k-ANONYMISIERUNG

Der Schutz sensibler Daten ist in der Gesetzgebung vieler Länder verankert. In Deutschland ist der Schutz sensibler Daten u. a. im Bundesdatenschutzgesetz (BDSG) und im Bundesstatistikgesetz (BStatG) festgeschrieben [Kar08; Deu09; Deu07]. Die gesetzliche Entsprechung der U. S. A. findet sich im Health Insurance Portability and Accountability Act (HIPAA) [Fed02].

Im Gegensatz zum BDSG und BStatG ist die Art der Anonymisierung im HIPAA genau beschrieben. Der HIPAA beschreibt mehrere Stufen des Datenschutzes [MBM11, S. 1]. Die „HIPAA Privacy Rule" fordert eine De-Identifikation durch das Entfernen von bestimmten IDs [MBM11, S. 1]. Die „HIPAA Safe Harbour Rules" erfordern zusätzlich die Generalisierung einiger als QIDs angesehenen Merkmale [CT13, S. 4].

Frühe Algorithmen zur De-Identifikation von Mikrodaten, wie das `Scrub`-System von Sweeney [Swe96], beschränkten sich auf die „HIPAA Privacy Rule". Das μ-Argus System des niederländischen Statistikers Hundepool anonymisiert die Mikrodaten durch Generalisierung sämtlicher einzigartiger 2- und 3-Kombinationen von Attributwerten [HW97].

Diese Ansätze sind auf einen sehr speziellen Einsatzzweck zugeschnitten [Swe01, S. 80]. Dadurch entstand der Bedarf eines generellen Konzepts.

Ein solches wurde von Sweeney durch das Konzept der *k*-Anonymisierung[1] geschaffen [Swe02b].

Definition 25 (*k*-Anonymisierung)**.** *Sei $T(A_1, \ldots, A_n)$ eine Tabelle und QI_T der mit ihr assoziierte Quasi-Identifikator. Wir sagen T genügt der k-Anonymisierung oder T ist k-anonym genau dann, wenn für jedes Tupel $t_i \in T$ gilt: $t_i[QI_T]$ ist ununterscheidbar zu mindestens $k-1$ weiteren Tupeln $t_j[QI_T] \in T$ für $1 \leq i, j \leq |T|$.*

Korollar 2. *Trivialer Weise ist jede Tabelle mindestens 1-anonym. Wir fordern daher für die k-Anonymisierung ein $k \geq 2$.*

Die Definition der *k*-Anonymisierung induziert eine Klassenbildung bezüglich der QIDs. Sämtliche Tupel einer Klasse sind ununterscheidbar bezüglich der Attributwerte ihrer QIDs. Wir sprechen in diesem Fall von *Äquivalenzklassen* oder *k-anonymen Gruppen*.

Durch die Definition der *k*-Anonymisierung wird eine Anforderung an eine Tabelle vor deren Veröffentlichung gestellt. Zur Erfüllung dieser Anforderung existieren verschiedene Vorgehensweisen [LDR05, S. 51].

Generell kann zwischen Methoden der *Generalisierung* sowie Methoden der *Unterdrückung* von Attributwerten zum Erreichen einer *k*-Anonymisierung unterschieden werden. Wenngleich im vorherigen Kapitel die Unterdrückung als maximale Ebene einer Generalisierung eingeführt wurde, so existieren Verfahren, welche ausschließlich mit der Unterdrückung von Attributwerten arbeiten [HW97, S. 144]. Im Kontext dieser Diplomarbeit werden wir uns auf Methoden der Generalisierung beschränken.

Die Methoden der Generalisierung folgen den Prinzipien des *Global Recoding* oder *Local Recoding*.

Algorithmen, welche das Prinzip des Local Recoding anwenden, modifizieren einzelne Attributwerte ausgewählter Tupel um eine *k*-Anonymisierung herzustellen [Fun+10, S. 14:18]. Dem gegenüber steht das Prinzip des Global Recoding. Algorithmen welche dem Prinzip des Global Recoding folgen, modifizieren stets sämtliche Vorkommen eines Attributwerts der Tabelle [Fun+10, S. 14:18]. Eine extreme Ausprägung des Global Recoding stellt die Full-Domain Generalisierung dar [LDR05, S. 51]. Das Prinzip der Full-Domain Generalisierung verlangt stets die Generalisierung aller Instanzen eines Attributs. Das bedeutet, dass sämtliche Attributwerte eines Attributs sich zu jedem Zeitpunkt der Generalisierung auf einer Ebene des VGH befinden.

1 Der Name „*k*-Anonymisierung “ wurde von Pierangela Samarati erdacht [Swe02b, S. 13].

Es erscheint sinnvoll, ausschließlich Algorithmen des Local Recoding zu verwenden. Durch diese kann eine *optimale* Generalisierung der Attributwerte erfolgen [De +11, S. 7]. Der Terminus der Optimalität bezieht sich entweder auf die benötigte Anzahl von Generalisierungen oder auf die *Utility* der Tabelle. Die Utility beschreibt den Erhalt von Information in der Tabelle und kann mit sogenannten *Utility-Metriken* gemessen werden [Swe02a, S. 9]. Ein Beispiel einer solchen Metrik stellt die Precision Metrik von Sweeney [Swe02a, S. 9] dar. Eine Tabelle besitzt eine höhere Utility, je mehr unveränderte Attributwerte sie besitzt [Swe02a, S. 9]. Meyerson et al. sowie Aggarwal et al. zeigten, dass das Problem des Erreichens einer diesbezüglich optimalen k-Anonymisierung mithilfe des Local Recoding NP-vollständig ist [MW04; Agg+05].

Aufgrund der NP-Vollständigkeit verwenden viele Algorithmen ein Prinzip, welches der Full Domain Generalisierung folgt.

Nachfolgend wollen wir uns mit dem Datafly-Algorithmus (DF-Algorithmus) beschäftigen. Dieser Algorithmus ist aufgrund seiner Einfachheit geeignet, den Vorgang der Generalisierung einer Tabelle nachzuvollziehen. Der DF-Algorithmus gehört zur Klasse der *Bottom-Up*-Algorithmen. Dies bedeutet, dass er spezifische Attributwerte der untersten Ebene des VGH sukzessive in weniger spezifische Attributwerte höherer Ebenen transformiert. Eine andere Klasse von Algorithmen der Generalisierung und Unterdrückung, stellen die *Top-Down*-Algorithmen dar. Diese überführen zunächst sämtliche Attributwerte der Tabelle in die Domäne der höchsten Ebene des VGH. Anschließend führen sie ein *Spezialisierung* der Attributwerte durch, bis eine weitere Spezialisierung die k-Anonymisierung verletzen würde.

4.1.1 *Datafly*

Einer der ersten heuristischen k-Anonymisierungs-Algorithmen ist der DF-Algorithmus von Sweeney. Es handelt sich um einen *Greedy*-Algorithmus, welcher nach dem Prinzip der Generalisierung und Unterdrückung arbeitet.

In der Folge ist das Ergebnis des Algorithmus nicht optimal hinsichtlich der minimalen Anzahl an anonymisierten Werten. Wir wollen uns nun eingehender dem Algorithmus widmen, da er durch seine einfache Struktur einen guten Zugang zu der Thematik der k-Anonymisierung gestattet. Er findet außerdem Verwendung im späteren Verlauf dieser Arbeit, so dass sich eine tiefer gehende Betrachtung lohnt.

Eine Übersicht des originalen Algorithmus ist im Pseudocode nach Sweeney [Swe02a, S. 13] ist in Algorithmus 4.1 dargestellt.

Startpunkt des Algorithmus ist seine erwartete Eingabe. Er erhält unveränderte tabellarische Mikrodaten sowie Informationen über die quasi-identifizierenden

Attribute. Wie wir im vorherigen Abschnitt gesehen haben, sind die Attribute des Typs QID für das Erreichen der k-Anonymisierung von besonderer Bedeutung. Der DF-Algorithmus erhält daher eine Liste von QIDs QT_T aufgrund derer eine k-Anonymisierung für ein gegebenes k erfolgt. Um diese nach dem Prinzip der Generalisierung und Unterdrückung durchführen zu können, erhält der Algorithmus weiterhin einen VGH für jedes quasi-identifizierende Attribut $A_i \in QI_T$.

Algorithmus 4.1 : Datafly

Eingabe : T, QI_T, k, VGH_{A_i} für jedes $A_i \in QI_T$
Ausgabe : k-anonyme Tabelle T_k

1 freq $\leftarrow$ Liste jeder Sequenz $t[QI_T]$ von quasi-identifizierenden Attributwerten und deren absoluten Häufigkeit $n(t[QI_T])$ in T
2 **while** $\exists t[QI_T] \in$ freq *mit* $n(t[QI_T]) < k$ **and** $\texttt{COUNT}(n(t[QI_T]) < k)) > k$ **do**
3 $A_j \leftarrow$ Attribut aus $\Pi_{QI_T}(T)$ mit den meisten verschiedenen Werten
4 freq $\leftarrow$ generalisiere alle Werte für A_j in freq
5 **end**
6 freq $\leftarrow$ unterdrücke Sequenzen in freq für die gilt: $n(t[QI_T]) < k$
7 freq $\leftarrow$ stelle k für unterdrückte Sequenzen in freq sicher
8 $T_k \leftarrow$ erstelle Tabelle aus freq
9 **return** T_k

Grundlage des Algorithmus ist ein assoziatives Datenfeld, welches den auftretenden QIDs ihre Häufigkeiten zuordnet. Dabei repräsentiert jede Sequenz eines QID ein oder mehrere Tupel der Tabelle und bildet damit die Äquivalenzklasse des QID. Die Konstruktion des Feldes geschieht in Zeile 1. Das heuristische Vorgehen des Algorithmus wird in Zeile 2 ersichtlich. Der Algorithmus fährt fort solange Äquivalenzklassen existieren, welche weniger als k Tupel enthalten. Zusätzlich muss die Anzahl der Tupel, welche nicht in k-anonymen Äquivalenzklassen organisiert sind, größer als k sein. Die Auswahl des zu generalisierenden Attributs erfolgt auf Grundlage der Variabilitäten der Attributwerte. Die Generalisierung des erwählten Attributs erfolgt gemäß der Full-Domain Generalisierung. Entsprechend wird die Generalisierung auf alle Werte in dem Datenfeld übertragen. Sollten nach Erreichen der Abbruchbedingung in Zeile 2 Äquivalenzklassen mit einem geringeren Umfang als k auftreten, so werden die Werte des QID für diese unterdrückt. Da in der Folge Tupel veröffentlicht werden, welche keinen QID aufweisen, wird in Zeile 7 auch für diese eine k-Anonymisierung gewährleistet. Dies kann beispielsweise durch Unterdrücken weiterer Äquivalenzklassen geschehen.

Das Ergebnis ist eine k-Anonymisierung bzgl. der QIDs. Diese wird auf die ursprüngliche Tabelle übertragen und als k-anonyme Tabelle T_K zurückgegeben .

Wir werden die Funktionsweise des Algorithmus nun näher betrachten, indem wir beispielhaft eine 3-Anonymisierung der Tabelle 3.2 vornehmen.

Beispiel 4. Seien die Eingaben mit T = Tabelle 3.2, $k = 3$ sowie $QI_T = \{\texttt{PLZ, Alter, Geschlecht}\}$ und den zugehörigen VGHs aus Abbildung 3.1b, 3.1c und 3.1d gegeben. Durch die Einzigartigkeit der Ausprägungen der Attribute `Alter` und `PLZ` besitzt das Datenfeld `freq` aus Algorithmus 4.1 neun Äquivalenzklassen – eine für jedes Tupel. Im ersten Generalisierungsschritt würde demnach eines der beiden Attribute auswählt. Wir wählen das Attribut `PLZ` und generalisieren die Werte gemäß VGH 3.1b. Das Ergebnis der Generalisierung sowie die Abfolge der weiteren Generalisierungen ist in den Tabellen 4.1a–4.1d dargestellt. Das Ergebnis der 3-anonymen Veröffentlichung findet sich in Tabelle 4.2a. Der Übersichtlichkeit halber wurden alle SAs bis auf das Attribut `Krankheit` ausgeblendet. Es bleibt darauf hinzuweisen, dass es sich bei diesem Ergebnis um eine mögliche k-Anonymisierung handelt. Bei Gleichheit des auszuwählenden Attributs in Zeile 2 des Algorithmus entschieden wir uns im ersten Schritt des Algorithmus für `PLZ`. Eine andere Möglichkeit hätte in dem Attribut `Alter` bestanden. Eine 3-Anonymisierung für diesen Fall findet sich in Tabelle 4.2b. Für dieses Beispiel wurde das SA `CRP` hinzugenommen.

4.1.2 *Schwächen der k-Anonymisierung*

Im vorherigen Abschnitt konnten wir das Erstellen einer k-Anonymisierung anhand des DF-Algorithmus nachvollziehen. Gemäß Definition 25 ist die Privatsphäre der in Tabelle 3.2 enthaltenen Individuen gegen Linking geschützt.

Wir wollen diese Aussage genauer überprüfen. Nach Li et al. [LLV07] muss bei Angriffen auf eine Anonymisierung zwischen zwei Arten der Informationspreisgabe unterscheiden werden:

1. Identity Disclosure sowie
2. Attribute Disclosure.

Tritt eine Identity Disclosure auf, so ist es einem Angreifer gelungen, einen Eintrag der veröffentlichten und anonymisierten Tabelle einem Individuum zuzuordnen – beispielsweise durch Linking. Im Falle der Attribute Disclosure ist es dem Angreifer möglich, einem Individuum einen oder mehrere Attributwerte der Veröffentlichung zuzuordnen. Es werden also zusätzliche Informationen über ein Individuum

ID	PLZ	ALTER	GESCHLECHT
1	4767*	29	46,XY
2	4760*	22	46,XY
3	4767*	27	46,XY
4	4790*	43	45,X0
5	4790*	49	46,XX
6	4790*	47	46,XX
7	4760*	30	47,XXY
8	4767*	36	46,XY
9	4760*	32	46,XY

(a) Generalisierung des Attributs `PLZ`

ID	PLZ	ALTER	GESCHLECHT
1	4767*	[20-29]	46,XY
2	4760*	[20-29]	46,XY
3	4767*	[20-29]	46,XY
4	4790*	[40-49]	45,X0
5	4790*	[40-49]	46,XX
6	4790*	[40-49]	46,XX
7	4760*	[30-39]	47,XXY
8	4767*	[30-39]	46,XY
9	4760*	[30-39]	46,XY

(b) Generalisierung des Attributs `Alter`

ID	PLZ	ALTER	GESCHLECHT
1	4767*	[20-29]	männlich
2	4760*	[20-29]	männlich
3	4767*	[20-29]	männlich
4	4790*	[40-49]	weiblich
5	4790*	[40-49]	weiblich
6	4790*	[40-49]	weiblich
7	4760*	[30-39]	männlich
8	4767*	[30-39]	männlich
9	4760*	[30-39]	männlich

(c) Generalisierung des Attributs `Geschlecht`

ID	PLZ	ALTER	GESCHLECHT
1	476**	[20-29]	männlich
2	476**	[20-29]	männlich
3	476**	[20-29]	männlich
4	479**	[40-49]	weiblich
5	479**	[40-49]	weiblich
6	479**	[40-49]	weiblich
7	476**	[30-39]	männlich
8	476**	[30-39]	männlich
9	476**	[30-39]	männlich

(d) Generalisierung des Attributs `PLZ`

Tabelle 4.1: Abfolge der Generalisierungen des Datafly-Algorithmus

erlangt, ohne dass dieses eindeutig mit einem Eintrag aus der Tabelle in Verbindung gebracht werden kann.

Die *k*-Anonymisierung einer Tabelle schützt gegen Identity Disclosure, nicht jedoch gegen Attribute Disclosure [LLV07, S. 2].

Dies begründet sich in dem Entwurf der *k*-Anonymisierung. Dieser betrachtet QIDs als die einzige Gefahrenquelle für den Schutz einer Tabelle [De +11, S. 3]. Es existieren zwei prominente Angriffe auf die *k*-Anonymisierung, welche eine Attribute Disclosure herbeiführen. Zum einen die *Homogeneity Attack*, zum anderen die *Background Knowledge Attack* [Mac+07, S. 3ff].

Wir wollen uns diese Begriffe anhand eines Beispiels verdeutlichen. Betrachten wir zunächst die Homogeneity Attack.

Beispiel 5 (Homogeneity Attack). Nehmen wir an, es würde eine 3-Anonymisierung der Tabelle 3.2 erfolgen. Das Resultat entspräche der Tabelle 4.2b. Nehmen wir wei-

ID	PLZ	ALTER	GESCHL.	KRANKHEIT
1	476**	[20-29]	männlich	Magengeschw.
2	476**	[20-29]	männlich	Gastritis
3	476**	[20-29]	männlich	Magenkrebs
4	479**	[40-49]	weiblich	Gastritis
5	479**	[40-49]	weiblich	Grippe
6	479**	[40-49]	weiblich	Bronchitis
7	476**	[30-39]	männlich	Bronchitis
8	476**	[30-39]	männlich	Magenkrebs
9	476**	[30-39]	männlich	Lungenentz.

(a) Mögliche 3-Anonymisierung

ID	PLZ	ALTER	GESCHL.	KRANKHEIT	CRP
1	4767*	[20-39]	männlich	Magengeschw.	erhöht
8	4767*	[20-39]	männlich	Magenkrebs	erhöht
3	4767*	[20-39]	männlich	Magenkrebs	erhöht
4	4790*	[40-59]	weiblich	Gastritis	erhöht
5	4790*	[40-59]	weiblich	Grippe	normal
6	4790*	[40-59]	weiblich	Bronchitis	leicht erhöht
7	4760*	[20-39]	männlich	Bronchitis	leicht erhöht
2	4760*	[20-39]	männlich	Gastritis	stark erhöht
9	4760*	[20-39]	männlich	Lungenentz.	normal

(b) Alternative 3-Anonymisierung

Tabelle 4.2: Resultate einer 3-Anonymisierung

ter an, ein *Angreifer* begehre Informationen über die Person „Heinrich“ bezüglich des Attributs CRP. Ein Angreifer bezeichnet in diesem Kontext eine nicht näher spezifizierte Institution, deren Ziel darin besteht die Anonymisierung einer Tabelle zu brechen, d. h. die Daten zu re-identifizieren. Die 3-Anonymisierung aus Tabelle 4.2b verhindert eine direkte Zuordnung der Person zu einem SA. Verfügt der Angreifer jedoch über das Wissen, dass es sich bei „Heinrich“ um einen Mann Mitte 30, aus dem Postleitzahlbereich „4767*“ handelt, so erfolgt eine Attribute Disclosure. Dies begründet sich in der Verschiedenheit der Attributwerte des SA CRP. Die Tupel mit der ID 1,8 und 3 bilden bezüglich der Werte „4767*, [20-29], männlich“ eine Äquivalenzklasse. In dieser Äquivalenzklasse findet sich ausschließlich der Wert „erhöht“ für das SA CRP. Der Angreifer benötigt daher keine exakte Zuordnung des SA zu der ID um zu folgern, dass das Attribut CRP von „Heinrich“ den Wert „erhöht“ besitzt.

Ein weiteres Beispiel soll uns die Intuition hinter dem Begriff der Background Knowledge Attack verdeutlichen.

Beispiel 6 (Background Knowledge Attack). Erneut wollen wir eine 3-Anonymisierung der Tabelle 3.2 mit dem Resultat der Tabelle 4.2b betrachten. Abermals begehre der Angreifer Informationen über die Person „Heinrich“ bezüglich des Attributs `Krankheit`. Bezogen auf die erste Äquivalenzklasse mit den IDs {1,8,3} beobachten wir eine Wahrscheinlichkeit von 66 %, dass „Heinrich“ an einer Magenkrebs-Erkrankung leidet. Die Wahrscheinlichkeit an Magengeschwüren erkrankt zu sein beträgt hingegen 33 %. Verfügt der Angreifer jedoch über das Vorwissen, dass in einer vorhergegangenen Untersuchung ein Magengeschwür-Leiden ausgeschlossen wurde, so kann er „Heinrich“ ein Magenkrebs-Leiden zu 100 % zuordnen.

Um diesen Angriffen entgegenzuwirken, entwickelten Machanavajjhala et al. das Konzept der ℓ-Diversity [Mac+07]. Wir werden uns diesem im nächsten Abschnitt widmen.

4.2 ℓ-DIVERSITY

Das Prinzip der ℓ-Diversity ist das erste Anonymisierungskonzept, welches das Hintergrundwissen eines Angreifers über die Anonymisierung modelliert [Veno8, S. 90f].

Die ℓ-Diversity wird als Ergänzung einer k-Anonymisierung verstanden [Mac+07, S. 4]. Wir werden sie gemäß Machanavajjhala et al. [Mac+07, S. 16] definieren:

Definition 26 (ℓ-Diversity)**.** *Gegeben eine k-anonyme Tabelle $T_k(A_1, \ldots, A_n)$ mit Attributen $A_1, \ldots, A_n$. Bezeichne SA_{T_k} ein sensibles Attribut aus T_k. Eine Äquivalenzklasse $E \in T_k$ genügt der ℓ-Diversity genau dann, wenn sie mindestens ℓ „wohlunterschiedene Werte" für das sensible Attribut SA_{T_k} aufweist.*

Gleichsam genügt die Tabelle der ℓ-Diversity genau dann, wenn jede Äquivalenzklasse ℓ-divers ist.

Die Autoren verstehen die ℓ-Diversity als *Prinzip*, da die Definition 26 den Begriff der „Wohlunterschiedenheit" nicht genauer spezifiziert. Nach Machanavajjhala et al. bildet das Prinzip der ℓ-Diversity einen Oberbegriff für Konzepte zur Messung der Homogenität von Attributwerten einer Äquivalenzklasse [Mac+07, S. 17].

Bereits vor dessen Veröffentlichung beschäftigten sich Arbeiten mit dem mangelnden Schutz de-anonymisierter Daten gegen Identity Disclosure. Noch vor der Formalisierung der k-Anonymisierung untersuchten Øhrn et al. das Problem homogener Attributwerte in Äquivalenzklassen [ØO99, S. 243].

Machanavajjhala et al. übertrugen die Erkenntnisse von Øhrn et al. in den Kontext der k-Anonymisierung und deren Schutz gegen die Identity Disclosure und Background Knowledge Attack. Die resultierende Instantiierung der ℓ-Diversity – die Entropy ℓ-Diversity – ist nach Machanavajjhala et al. [Mac+07, S. 17] wie folgt definiert:

Definition 27 (Entropy ℓ-Diversity)**.** *Gegeben eine k-anonyme Tabelle $T_k(A_1, \ldots, A_n)$ mit Attributen $A_1, \ldots, A_n$. Bezeichne E eine Äquivalenzklasse aus T_k. Sei weiter $SA_{T_k} \in \{A_1, \ldots, A_n\}$ ein sensibles Attribut. Dessen Domäne umfasse die Attributwerte $\{x_1, \ldots, x_m\}, 1 \leq m \leq |E|$ mit der dazugehörigen Häufigkeitsverteilung X_E.*

Wir sagen, E genügt der Entropy ℓ-Diversity bezüglich des sensiblen Attributs SA_{T_k} genau dann, wenn für jedes Tupel $t_i \in E$:

$$-\sum_i f_{X_E}\left(t_i[SA_{T_k}]\right) \log_2\left(f_{X_E}\left(t_i[SA_{T_k}]\right)\right) \geq \log_2(\ell) \tag{4.1}$$

Die Definition nutzt die Eigenschaft, dass die Entropie einer Menge von Attributwerten steigt je näher die Häufigkeitsverteilung der Gleichverteilung kommt (vgl. Abschnitt 2.2.3).

Eine naivere Instanz der ℓ-Diversity wird als *Distinct ℓ-Diversity* bezeichnet. Das Konzept der *Distinct ℓ-Diversity* verlangt für jedes Attribut einer k-anonymen Tabelle mindestens ℓ verschiedene Attributwerte pro Äquivalenzklasse. Diese Definition ergibt sich als Folge der Definition 27 [Mac+07, S. 17]. Eine äquivalente Formulierung findet sich in dem Konzept der *p-senitive k-Anonymity* von Truta et al. [TV06].

Es existieren weitere Instanzen dieses Konzepts, auf die wir im Kontext dieser Arbeit nicht eingehen wollen (vgl. Machanavajjhala et al. [Mac+07, S. 18ff]).

Wir werden nachfolgend die Entropy ℓ-Diversity der k-anonymen Tabelle 4.2b bestimmen. Dies soll uns ein Gefühl für den Umgang mit Äquivalenzklassen und der Häufigkeitsverteilung der Attributwerte ihrer SAs verschaffen.

Beispiel 7 (Entropy ℓ-Diversity). In Tabelle 4.2b erkennen wir drei Äquivalenzklassen. Diese sind durch die IDs $E_1 = \{1,8,3\}$, $E_2 = \{4,5,6\}$, und $E_3 = \{7,2,9\}$ gekennzeichnet. Es bezeichnen X_{E_1} X_{E_2} und X_{E_3} die Häufigkeitsverteilungen des sensiblen Attributs `Krankheit`. Dann gilt:

$$\begin{aligned} -\sum_i f_{X_{E_1}}(x_i)\log_2\left(f_{X_{E_1}}(x_i)\right) &= -f(\texttt{Magengeschw.})\log_2\left(f(\texttt{Magengesch.})\right) + \\ &\quad + (-2)\cdot f(\texttt{Magenkrebs})\log_2\left(f(\texttt{Magenkrebs})\right) \\ &= -\frac{1}{3}\cdot\log_2\left(\frac{1}{3}\right) + -\frac{2}{3}\cdot\log_2\left(\frac{2}{3}\right) \\ &\approx 0,528 + 0,390 = 0,918 \end{aligned} \tag{4.2}$$

Durch die Umkehrfunktion des Logarithmus gilt für die Äquivalenzklasse E_1 : $\log_2(\ell) \leq 0,918 \iff \ell \leq 2^{0,918} = 1,889$. Es folgen die Berechnungen für E_2 und E_3 :

$$\begin{aligned} -\sum_i f_{X_{E_2}}(x_i)\log_2\left(f_{X_{E_2}}(x_i)\right) &= 3\cdot -\frac{1}{3}\cdot\log_2\left(\frac{1}{3}\right) \\ &\approx 3\cdot 0,528 = 1,585 \overset{2^{1,585}}{\Longrightarrow} \text{3-divers} \end{aligned} \tag{4.3}$$

$$\begin{aligned} -\sum_i f_{X_{E_3}}(x_i)\log_2\left(f_{X_{E_3}}(x_i)\right) &= 3\cdot -\frac{1}{3}\cdot\log_2\left(\frac{1}{3}\right) \\ &\approx 3\cdot 0,528 = 1,585 \overset{2^{1,585}}{\Longrightarrow} \text{3-divers} \end{aligned} \tag{4.4}$$

Wir stellen fest, dass Tabelle 4.2b Entropy 1,889-Diverse ist, da dies den kleinsten Wert der Entropie für alle Äquivalenzklassen darstellt.

Wir werden uns in diesem Abschnitt nicht näher mit Algorithmen zum Erreichen einer ℓ-Diversity beschäftigen. Wenngleich die ℓ-Diversity eine wichtige Erweiterung der k-Anonymisierung darstellt, so dient sie uns im Kontext dieser Arbeit lediglich

als Überleitung zu einem mächtigeren Datenschutzkonzept: Der t-Closeness. Zu diesem Zweck werden wir im nächsten Abschnitt Schwachstellen der ℓ-Diversity aufzeigen und uns anschließend dem Konzept der t-Closeness widmen.

4.2.1 *Schwächen der ℓ-Diversity*

Im letzten Abschnitt haben wir die Entropy ℓ-Diversity für Tabelle 4.2b berechnet. Wir haben uns bislang nicht mit der Frage beschäftigt, wie hoch der Schutz der sensiblen Daten ist, welcher sich durch die 3-Anonymisierung mit der Entropy ℓ-Diversity von $1,889$ ergibt.

Die Autoren des Konzepts der t-Closeness konnten hinsichtlich des Schutzes der ℓ-Diversity zwei Angriffe identifizieren. Zum einen die *Skewness Attack* und zum anderen die *Similarity Attack* [LLV07, S. 3].

Betrachten wir zunächst die Skewness Attack anhand eines Beispiels:

Beispiel 8 (Skewness Attack). Widmen wir uns abermals der 3-Anonymisierung aus Tabelle 4.2b. Wir gehen erneut von einem Angreifer aus, welcher Informationen über die Person „Heinrich“ bezüglich des Attributs `Krankheit` zu erlangen versucht. Erneut ist dem Angreifer bekannt, dass es sich bei „Heinrich“ um einen Mann Mitte 30, aus dem Postleitzahlbereich „4767*“ handelt. Er kann folglich der ersten Äquivalenzklasse zugeordnet werden. Demzufolge beträgt die Wahrscheinlichkeit, dass „Heinrich“ an einer Magenkrebs-Erkrankung leidet 66 %. Verglichen mit der auf die gesamte Tabelle bezogenen Wahrscheinlichkeit an Magenkrebs zu erkranken von 22 %, stellt dies eine deutliche Erhöhung dar. Diese Erhöhung könnte den Angreifer dazu verleiten „Heinrich“ Magenkrebs zu attestieren. Dies liegt in der relativen Häufigkeit der Attributwerte begründet. Die Häufigkeitsverteilung des Attributs `Krankheit` ist *schief*. Dies bedeutet, dass einzelne Attributwerte in einigen Äquivalenzklassen häufiger auftreten als in anderen.

Als weiteren Schwachpunkt der ℓ-Diversity konnten Li et al. die Similarity Attack ausmachen. Auch diese wollen wir anhand eines Beispiels betrachten.

Beispiel 9 (Similarity Attack). In diesem Fall wollen wir uns der Tabelle 4.2a zuwenden. Die 3-Anonymisierung erweist sich als Entropy 3-Diverse in Bezug auf das SA `Krankheit`. Gemäß Definition 27 bietet diese k-Anonymisierung die maximale Sicherheit der Daten gegen Attribute Disclosure. Dies begründet sich mit der Gleichverteilung der Attributwerte bezogen auf jede Äquivalenzklasse. Auch in diesem Fall kann ein Angreifer, welcher „Heinrich“ in der ersten Äquivalenzklasse vermutet, Schlüsse über seine Erkrankung ziehen. Ungeachtet dessen, unter welcher Krankheit „Heinrich“ leidet, steht fest, dass er wegen einer Erkrankung des Magens behandelt

wurde. Auch dies kann als Verletzung der Privatsphäre angesehen werden. Die Similarity Attack begründet sich in der Tatsache, dass die ℓ-Diversity eine syntaktische Anforderung an die Attributwerte stellt – die Verschiedenheit der Ausprägungen – jedoch keine Vorgaben bezüglich der Semantik der Attributwerte trifft [De +11, S. 10].

Wir konnten sehen, dass auch das Konzept ℓ-Diversity Schwachstellen aufweist. Im nächsten Abschnitt werden wir uns dem bedeutenden Konzept der t-Closeness widmen, welche die Semantik der Attributwert in die Berechnung der Güte einer Anonymisierung einbezieht.

4.3 t-CLOSENESS

In Abschnitt 4.1.2 wurden die Identity Disclosure sowie die Attribute Disclosure als mögliche Gefahrenquellen für die Re-Identifikation von Individuen in einer Veröffentlichung behandelt. Wie in den Abschnitten 4.1.2 und 4.2.1 gezeigt, schützen k-Anonymisierung und ℓ-Diversity vor Identity Disclosure, jedoch nicht vor Attribute Disclosure. Um diesem Umstand entgegenzuwirken, entwickelten Li et al. ein Verfahren, welches in der Literatur als eine Erweiterung der ℓ-Diversity beschrieben wird [AYE08, S. 27]: Das Konzept der t-Closeness.

4.3.1 *Definition der t-Closeness*

Das Konzept der *Nähe* von Attributwerten, als wichtigem Schutz gegen Attribute Disclosure, wurde von den vorher genannten Metriken außer Acht gelassen. So erreichen diese zwar eine Nicht-Unterscheidbarkeit der sensiblen Attribute – und unter diesen auch eine gewisse Vielfalt – sie bieten jedoch sehr unterschiedliche Grade des Schutzes der Privatsphäre [LLV07]. Ausgehend von den in der Originalveröffentlichung zum Prinzip der t-Closeness genannten Beispielen, klassifizieren De Capitani di Vimercati et al. drei Angriffspunkte auf k-Anonymisierung und ℓ-Diversity. Diese Angriffspunkte entstehen durch die Vernachlässigung der Semantik der Attributwerte und können somit zu Attribute Disclosure führen:

1. Die Häufigkeitsverteilung der Attributwerte
2. Mögliche semantische Beziehungen zwischen den Attributwerten
3. Unterschiedliche Vertraulichkeitsgrade der Attributwerte

In Beispiel 9 ist nicht die Häufigkeitsverteilung der Attributwerte (Punkt 1) der ausschlaggebende Faktor für die Verletzung der Privatsphäre, sondern die Tatsache,

dass alle sensiblen Attributwerte der fokussierten Äquivalenzklasse ein bestimmtes Teilgebiet der Medizin betreffen. Sie besitzen demnach eine ähnliche Semantik (Punkt 2). Punkt 3 nimmt eine besondere Stellung ein. Ausgehend von Beispiel 9 ließe sich zudem folgern, dass das Ziel des Angreifers eventuell unter Magenkrebs leidet, was eine schwerwiegendere Diagnose als die vorliegenden Erkrankungen der Atemwegsorgane darstellt.

Betrachten wir Tabelle 4.2 als Ganzes. Es zeigt sich, dass Erkrankungen des Verdauungstrakts sowie der Lunge im Verhältnis 4 : 5 vorliegen. Es zeigt sich demnach keine eindeutige Häufung der einen Erkrankung gegenüber der anderen. Li et al. stellten fest, dass sich das Wissen des Angreifers durch Kenntnis der Veröffentlichung verändert.

Ausgehend von dieser Beobachtung geben Li et al. ein intuitives Maß für Privatsphäre:

> „Privatsphäre ist der Informationsgewinn eines Beobachters. Es gilt: Der Angreifer hat eine vorhergehende, apriorische Annahme über den Wert des sensiblen Attributs eines Individuums. Nach Einsicht in die anonymisierten Informationen besitzt der Angreifer eine aposteriorische Annahme über eben diesen Wert. Die Differenz dieser beiden Werte wird als *Informationsgewinn* bezeichnet."
>
> — Definition der Privatsphäre nach Li et al. [LLV07, S. 4]

Diese Definition bezieht das Hintergrundwissen eines Angreifers in den Schutz der Privatsphäre ein. Li et al. folgern, dass eine Anonymisierung, den Erkenntnisgewinn zwischen Hintergrundwissen und Wissen durch die Veröffentlichung der Daten minimieren muss. Bei der Konzeption der k-Anonymisierung sowie ℓ-Diversity wurde dies nicht berücksichtigt. Für die Modellierung nach Li et al. wird dem Angreifer Wissen über die anonymisierte Version der Veröffentlichung unterstellt. In dieser sind sämtliche Werte der QIDs maximal anonymisiert, d. h. durch einen Wert aus der Domäne des obersten Knotens des DGH ersetzt[2]. Aus dieser Veröffentlichung erhält der Angreifer das Wissen über die Häufigkeitsverteilung der sensiblen Attribute. Es bleibt zu betonen, dass dieser Schritt ein hypothetischer ist. Er dient allein dazu, um den Wissenszuwachs des Angreifers durch die Veröffentlichung der Tabelle beschreiben zu können.

Durch die Veröffentlichung der Tabelle und des Wissens über die quasi-identifizierenden Attribute seines Ziels, erhält der Angreifer die Häufigkeitsverteilung der sensiblen Attribute der Äquivalenzklasse, in welcher sich sein Ziel befindet.

2 Voraussetzung hierfür ist, dass auch die maximale Generalisierung eine valide k-Anonymisierung, in Bezug auf ein gegebenes k, darstellt.

Ausgehend von der vorherigen Überlegung konstatieren Li et al. [LLV07, S. 4]:

> „In some sense, the larger the difference between [the most general publication] and [the final publication] is, the more valuable the data is."

Große Unterschiede zwischen der „most general publication" – in Definition 4.3.1 als apriori-Wissen bezeichnet – sowie der „final publication" – dem aposteriorischen Wissen aus Definition 4.3.1 – ermöglichen folgerichtig einen starken Informationsgewinn. Daher erscheint es sinnvoll, die Distanz zwischen den beiden Verteilungen zu minimieren. Sie definieren daher das Konzept der t-Closeness wie folgt:

Definition 28 (t-Closeness)**.** *Eine Äquivalenzklasse besitzt t-Closeness, sobald der Abstand der Häufigkeitsverteilung jedes Wertes ihrer sensiblen Attribute zu der Häufigkeitsverteilung der Werte der sensiblen Attribute in der gesamten Veröffentlichung nicht mehr als einen Grenzwert t beträgt. Genügen alle Äquivalenzklassen der Veröffentlichung der t-Closeness, so besitzt die Veröffentlichung selbst t-Closeness.*

Bis zu diesem Punkt haben wir die Intuition hinter dem Konzept der t-Closeness eingeführt. Dabei wurde der Terminus *Nähe von Häufigkeitsverteilungen* benutzt (vgl. Seite 41), ohne diesen genauer erläutert zu haben. Dieses wird im folgenden Abschnitt ergänzt.

4.3.2 *Earth-Movers-Distance*

Der Begriff der Nähe verlangt eine Definition von *Abstand*, um die tatsächliche Nähe zwischen zwei Objekten messen zu können. Einen solchen Abstand werden wir in Anlehnung an Rubner et al. als *Ground-Distance* bezeichnen [RTG98]. Den Begriff des metrischen Raums sowie den der Metrik haben wir bereits in Kapitel 2 eingeführt. Diese sollen nun der Definition eines Abstandsmaßes zwischen zwei Verteilungen dienen. Anhand dessen werden wir uns den Begriff der semantischen Nähe erschließen.

4.3.2.1 *Ground-Distance*

Wir wissen bereits, dass Häufigkeitsverteilungen von Attributwerten als Säulendiagramm dargestellt werden können (siehe Abschnitt 2.1.4). Die Ground-Distance beschreibt ein Abstandsmaß zwischen den Säulen des Diagramms. Nach Li et al. muss dieses auf einem metrischen Raum definiert sein und Werte im Intervall $[0, 1]$ annehmen [LLV07, S. 5].

In der Literatur existieren zahlreiche Beispiele, welche sich mit der Berechnung des Abstandes zwischen zwei Häufigkeitsverteilungen beschäftigen. Sie haben ihre

	GRIPPE (x_1)	LUNGENENT. (x_2)	BRONCH. (x_3)	MAGENG. (x_4)	MAGENK. (x_5)	GASTR. (x_6)
V	1/9	1/9	2/9	1/9	2/9	2/9
E_1	0	0	0	1/3	1/3	1/3
E_2	1/3	0	1/3	0	0	1/3
E_3	0	1/3	1/3	0	1/3	0

Tabelle 4.3: relative Häufigkeit der sensiblen Attributwerte

Anwendung im Bereich der Multimedia-Datenbanken, lassen sich aber aufgrund der ähnlichen Darstellung des Problems – Bilddaten werden als Häufigkeitsverteilungen von Farbtönen dargestellt – auf den Bereich der Datenanonymisierung übertragen. Populäre Abstandsmaße sind laut Rubner et al. [RTG98] die durch die p-Normen induzierten Metriken. Diese werden auch als *Minkowski-Metriken* bezeichnet, deren prominenteste Ausprägungen in Form der *Manhattan-Distanz (MH-Distanz)* oder des *Euklidischen Abstands* bekannt sind [Sch06a, S. 170].

Eine andere Abstandsfunktion stellt die *Kullback-Leibler-Distanz (KL-Distanz)* dar, welche auch als *relative Entropie* bezeichnet wird. Sie ist jedoch keine Metrik gemäß Definition 1, da sie die Symmetrie- (M2) sowie Dreiecksungleichungs-Eigenschaft (M3) nicht erfüllt [Sch06a, S. 185ff]. Trotz dieser Einschränkung ist diese ein weit verbreitetes Maß, um den Abstand zweier Verteilungen zu bestimmen [Bez10; Sch06a].

Im Bereich der Datenanonymisierung fand die KL-Distanz bereits zur Quantifizierung der Utility von Datenanonymisierungen Verwendung [Mac+07]. So erwähnen auch Li et al. die KL-Distanz als Maß für den Abstand zweier Verteilungen und somit als Funktion zur Bestimmung einer Nähe. Die vorher genannten Abstandsfunktionen beziehen jedoch nicht die Semantik der Attributwerte in ihre Abstandsberechnung mit ein [LLV07]. Ein Rechenbeispiel, welches diesen Umstand für die KL-Distanz verdeutlicht, findet sich in Beispiel 20 aus Anhang B.

Dass die KL-Distanz unser Verständnis von Nähe nicht erfüllt, ergibt sich aus der Tatsache, dass diese zwar eine Ground-Distance zwischen einzelnen Elementen einer Verteilung definiert, der Abstand zwischen allen Elementen jedoch identisch ist. Li et al. benennen diese Eigenschaft als das Fehlen der *Semantic Awareness*. Sie definieren daher die „Desiderata for Designing a Distance Measure“ [LLV10, S. 949].

Die Earth-Movers-Distanz (EMD) ist ein Distanzmaß, welches das Kriterium der Semantic Awareness erfüllt.

4.3.2.2 *Die Earth-Movers-Distance als lineares Optimierungsproblem*

Ursprünglich von Rubner et al. [RTGoo] vorgeschlagen, um die Suche auf Bild-Datenbanken zu optimieren, wurde die EMD von Li et al. [LLV07] aufgenommen und in den Bereich der Datenanonymisierung übertragen. Die EMD ist eine Ausprägung des *Transportproblems*. Für den Kontext dieser Arbeit relevant formuliert, handelt es sich bei dem Transportproblem um die Lösung der Frage, wie der Aufwand für die Überführung einer Wahrscheinlichkeitsverteilung in eine andere minimiert werden kann. Intuitiv klar ist, dass dieser Aufwand geringer ist, je näher sich die beiden Verteilungen sind. In diesem Zusammenhang sei daran erinnert, dass es sich bei der Veröffentlichung von sensiblen Daten um die Veröffentlichung von Mikrodaten handelt. Die Werte der Attribute einer Tabelle bilden eine Häufigkeitsverteilung und auch die Werte einer Teilmenge der Veröffentlichung – wie etwa die Äquivalenzklassen – bilden eine Häufigkeitsverteilung über den in ihnen enthaltenen Attributwerten.

Das Transportproblem kann laut Rubner et al. durch Methoden der *linearen Optimierung* gelöst werden. Die lineare Optimierung beschäftigt sich mit dem Problem „eine lineare Funktion [. . .] unter Berücksichtigung linearer Gleichungen und Ungleichungen als Nebenbedingungen [. . .] zu maximieren oder zu minimieren“ [SM88, S. 10].

Für ein besseres Verständnis der EMD soll diese nun visualisiert werden. Das nachfolgende Beispiel wird uns anhand der Umformung eines Säulendiagramms die Arbeitsweise der EMD verdeutlichen, ohne auf die Details der EMD einzugehen. Anschließend werden wir Algorithmen zur Berechnung der EMD betrachten.

Beispiel 10. Wir betrachten die Tabelle 3.2 und die daraus abgeleitete 3-Anonymisierung aus Tabelle 4.2a. Aus diesen beiden Darstellungen lassen sich für die gesamte Tabelle sowie die in ihr enthaltenen Äquivalenzklassen Häufigkeitsverteilungen ablesen.

Wir beschränken uns auf die Äquivalenzklasse E_1, welche aus den Tupeln $\{1, 2, 3\}$ gebildet wird. Diese ist vereinfacht in Abbildung 4.1a dargestellt. Es gilt E_1 mit der gesamten Veröffentlichung zu vergleichen, welche wir als V bezeichnen wollen und welche auszugsweise in Abbildung 4.1b dargestellt ist. Die relativen Häufigkeiten der Attributwerte können wir Tabelle 4.3 entnehmen. Mithilfe der EMD wollen wir die Häufigkeitsverteilung X_{E_1} in die Häufigkeitsverteilung X_V überführen.

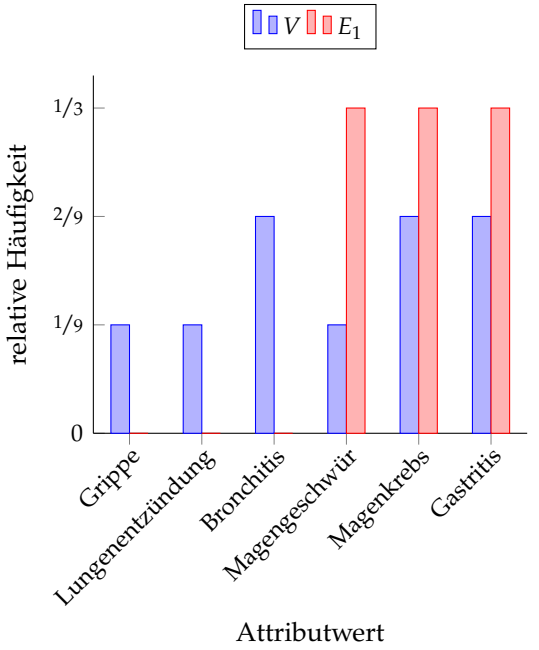
Bildlich gesprochen besteht X_{E_1} aus Erdhaufen, die der Größe ihrer Häufigkeit entsprechend über der Menge der möglichen Attributwerte (Merkmalsausprägungen) verteilt sind. Analog zu dieser Assoziation besteht X_V aus Erdlöchern, deren Tiefe den relativen Häufigkeiten über derselben Menge an Attributwerten entspricht. Die Überführung ist vollständig, sobald alle Erdlöcher mit der Erde aus den Erdhaufen gefüllt wurden. Übertragen wir das Bild der Erdhaufen und -Löcher auf ein Säu-

ID	PLZ	ALTER	KRANKHEIT
1	476**	[20-29]	Magengeschwür
2	476**	[20-29]	Gastritis
3	476**	[20-29]	Magenkrebs
...	...	...	...

(a) Äquivalenzklasse E_1

ID	PLZ	ALTER	KRANKHEIT
1	47677	29	Magengeschwür
2	47602	22	Gastritis
3	47678	27	Magenkrebs
4	47905	43	Gastritis
5	47909	49	Grippe
6	47906	47	Bronchitis
7	47605	30	Bronchitis
8	47673	36	Magenkrebs
9	47607	32	Lungenentzündung

(b) Gesamte Veröffentlichung V

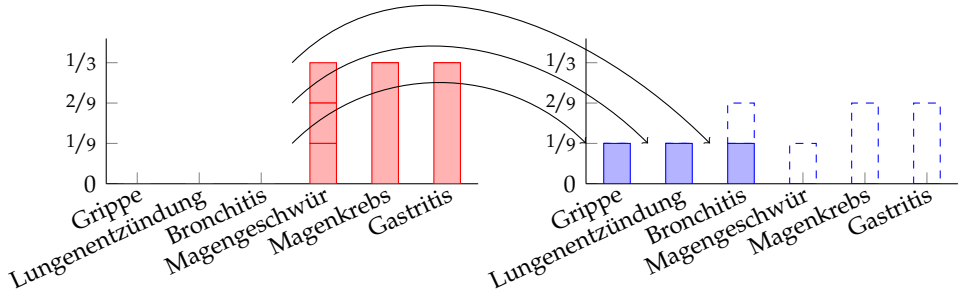


(c) Gegenüberstellung der Verteilungen von E_1 und V

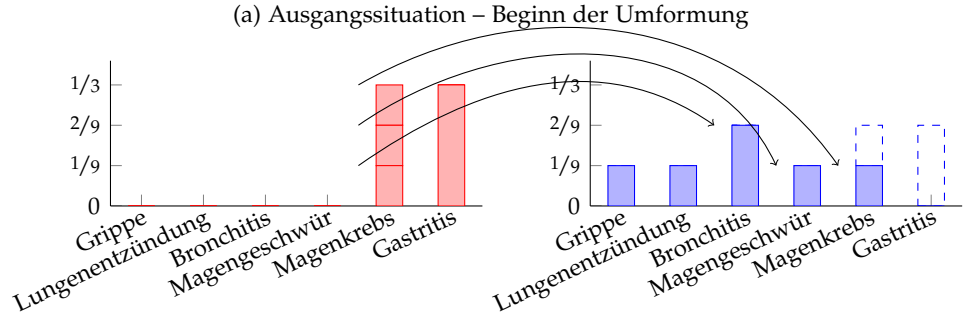
Abbildung 4.1: Häufigkeitsverteilungen von E_1 und V

lendiagramm, erhalten wir eine Darstellung, wie sie in Abbildung 4.1c zu sehen ist. Die verschiedenfarbigen Säulen stellen die jeweiligen Verteilungen der Attributwerte dar. Eine mögliche Abfolge an Bewegungen sowie das Ergebnis de Transformation ist den Abbildungen 4.2a, 4.1b und 4.2c zu entnehmen.

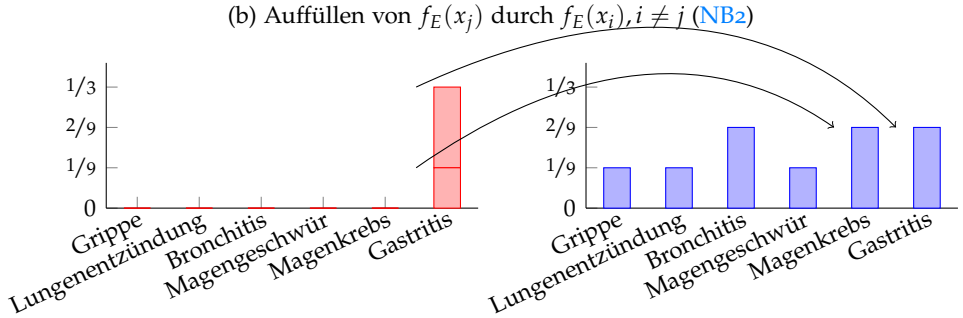
Dieses Beispiel diente der Verdeutlichung hat der Arbeitsweise der EMD. Nun wollen wir diese formalisieren. Seien X_E und X_V zwei Häufigkeitsverteilungen eines beliebigen sensiblen Attributs. X_E bezeichne die relativen Häufigkeiten der Attributwerte in einer belieben Äquivalenzklasse und X_V bezeichne die relativen Häufigkeiten der Attributwerte bezogen auf die gesamte Veröffentlichung. Die relativen Häufigkeiten von m Attributwerten seien durch $X_E = \{f_E(x_1), \ldots, f_V(x_m)\}$ sowie $X_V = \{f_V(x_1), \ldots, f_V(x_m)\}$ gegeben. Es seien weiter die Kosten für die Überführung eines Elements $x_i \in X_E$ in ein Element $x_j \in X_V$ mit $c(x_i, x_j)$ gegeben. Diese Kosten entsprechen der Ground-Distance (vgl. Abschnitt 4.3.2.1), welche die semantische Nähe zwischen den Attributwerten darstellt. Die zu transportierende Wahrscheinlichkeitsmasse, d. h. die relativen Häufigkeiten der Attributwerte, eines Elements $x_i \in X_E$ zu einem Element $x_j \in X_V$ sei durch $f(x_i, x_j)$ dargestellt. Ziel der EMD



(a) Ausgangssituation – Beginn der Umformung



(b) Auffüllen von $f_E(x_j)$ durch $f_E(x_i), i \neq j$ (NB2)



(c) Resultat der Transformation

Abbildung 4.2: Transformation von X_{E_1} nach X_V

ist es, eine Abfolge von Transportoperationen $F = \{f(x_{i_1}, x_{j_1}), f(x_{i_2}, x_{j_2}), \ldots\}$ von Wahrscheinlichkeitsmasse der einen Verteilung zur anderen zu finden, welche die Verteilung X_E in X_V überführt und deren Kosten minimal sind.

Nach [LLV07] und [RTG98] als Problem der linearen Optimierung formuliert bedeutet dies:

$$WORK(X_E, X_V, F) = \sum_{i=1}^{m} \sum_{j=1}^{m} c(x_i, x_j) f(x_i, x_j) \tag{4.5}$$

Bei Einhaltung der Nebenbedingungen:

$$f(x_i, x_j) \geq 0 \qquad 1 \leq i, j \leq m \qquad \text{(NB1)}$$

$$f_E(x_i) - \sum_{j=1}^{m} f(x_i, x_j) + \sum_{k=1}^{m} f(x_k, x_i) = f_V(x_i) \qquad 1 \leq i \leq m \qquad \text{(NB2)}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m} f(x_i, x_j) = \sum_{i=1}^{m} f_E(x_i) = \sum_{i=1}^{m} f_V(x_i) = 1 \qquad \text{(NB3)}$$

Nebenbedingung (NB1) definiert, dass nur Masse der Verteilung X_E zur Verteilung X_V transportiert werden darf und nicht andersherum. Für die Erläuterung von Nebenbedingung (NB2) wollen wir uns an dieser Stelle noch einmal vergegenwärtigen, dass wir die Verteilung X_E umformen. Das heißt, dass jedes Element an Position i in X_E in sein Pendant an der Stelle i in X_V transformiert werden muss, indem die relative Häufigkeit $f_E(x_i)$ des entsprechenden Attributwerts der relativen Häufigkeit $f_V(x_i)$ angepasst wird. Nebenbedingung (NB2) stellt dies sicher: Zum einen verlangt sie, dass sämtliche Masse eines Elementes $x_i \in X_E$ zur Gestaltung der Elemente $x_j \in X_V$ genutzt wird. Zum anderen stellt sie sicher, dass das Element $x_i \in X_V$ nötigenfalls aus anderen Elementen $x_k \in X_E, i \neq k, 1 \leq k \leq m$ gebildet wird, sollte die eigene Masse des Elements $x_i \in X_E$ nicht ausreichen. Nebenbedingung (NB3) festigt bereits Bekanntes: Die Summe der relativen Häufigkeiten beträgt 1 (vgl. Korollar 1), dies gilt für die beiden Verteilungen X_E und X_V und daher auch für die Summe des Masseflusses.

Ist ein Fluss F gefunden, der genau dies leistet, so ist die EMD definiert als:

$$d_{EMD}(X_E \| X_V) := WORK(X_E, X_V, F) = \sum_{i=1}^{m}\sum_{j=1}^{m} c(x_i, x_j) f(x_i, x_j) \qquad (4.6)$$

Nach Li et al. liegt das Ergebnis der Funktion $d_{EMD}(\cdot\|\cdot)$ zwischen 0 und 1. Dies ist der Faktor t, welcher den Grad der Anonymisierung der Daten bemisst. Der Fall $d_{EMD} = 0$ ist ein Spezialfall, welchen wir näher beleuchten wollen. So ist für einen Abstand von 0 eine Identität der beiden Verteilungen erforderlich. Wir halten fest:

Theorem 1 (Der Abstand identischer Verteilungen ist 0)**.** *Gegeben zwei Häufigkeitsverteilungen X_E und X_V. Dann gilt*

$$d_{EMD}(X_E \| X_V) = 0 \iff X_E = X_V \quad . \qquad (4.7)$$

Beweis. Wir führen den Beweis per Widerspruch und zeigen zunächst $d_{EMD}(X_E\|X_V) = 0 \implies X_E \neq X_V$. Sei also $d_{EMD}(X_E\|X_V) = 0$ und $X_E \neq X_V$. Da die Verteilungen ungleich sind, muss es mindestens ein x_i geben für das gilt:

$f_X(x_i) \neq f_V(x_i)$. Das bedeutet, dass es mindestens eine Massebewegung $f(x_i, x_j) > 0$ gegeben haben muss, welche Kosten $c(x_i, x_j) > 0$ verursacht hat. Folglich ist auch die Summe aller Flüsse von 0 verschieden. Dies steht im Widerspruch zur Voraussetzung.

Die Gegenrichtung verläuft analog. Wir setzen voraus, dass die Häufigkeitsverteilungen X_E und X_V identisch sind und dass daraus $d_{EMD}(X_E \| X_V) \neq 0$ folgt. Wenn nun die Verteilungen identisch sind, so gilt $\forall i : f_E(x_i) = f_V(x_i)$. Demnach ist die Ground-Distance $c(x_i, x_j) = 0$ für alle i, j. Daher muss ebenfalls die Summe der Produkte 0 sein, womit gilt: $d_{EMD}(X_E \| X_V) = 0$. Dies ist ein Widerspruch.

Zusammen folgt die Behauptung. □

Uns ist nun bekannt, dass die EMD eine Lösung für ein lineares Optimierungsproblem ist. Bisher fehlt jedoch ein Algorithmus, welcher die Lösung umsetzt. Rubner et al. verwendeten zur Berechnung der EMD bezüglich ihrer Problemstellung eine Variante des Simplex-Verfahrens von George Dantzig [RTGoo]. Laut Li et al. bietet dieses Verfahren jedoch keinen expliziten Lösungsweg für das gegebene Problem des Vergleichs zweier Wahrscheinlichkeitsverteilungen. Insbesondere berücksichtigt es nicht, dass für eine semantische Betrachtung der Attributwerte durch die EMD zwischen zwei Arten von Attributwert-Domänen unterschieden werden muss: quantitativen sowie kategorialen Attributwerten.

4.3.2.3 *Berechnung der Earth-Movers-Distance für quantitative Attributwerte*

Der Begriff des quantitativen Merkmals wurde bereits in Definition 6 auf Seite 10 formalisiert. Für den Fall, dass die Ground-Distance für quantitative sensible Attribute berechnet wird, schlagen Li et al. die *Ordered Distance* als Grundlage der Kostenfunktion der EMD vor. Somit gilt für die total geordneten Elemente der Attributwerte $\{x_1, \ldots, x_m\}$:

$$c(x_i, x_j) = ordered\_dist(x_i, x_j) = \frac{|i - j|}{m - 1} \tag{4.8}$$

Das von Li et al. eingeführte Verfahren der Ordered Distance bedient sich einer *Skalentransformation*, um die vormals metrisch skalierten Merkmale ordinal durch ihre *Rangzahl* zu repräsentieren. Dies bedeutet, dass den Werten entsprechend ihrer Ordnung eine natürliche Zahl zugewiesen wird, welche ihre Stellung zueinander beschreibt. Dividiert mit der Spannweite wird das Resultat auf das Einheitsintervall $[0, 1]$ skaliert. Dadurch ergibt sich eine Minimal-Distanz von 0 für identische Werte und eine Maximal-Distanz von 1 für die jeweiligen Extrema der Attributwertmenge. Diese Skalierung ist notwendig, um als Ergebnis der EMD einen Wert zwischen 0

und 1 und damit eine Grundlage für die Festlegung eines t für die Verwendung der t-Closeness zu besitzen [LLV07, S. 5].

Algorithmus 4.2 zeigt die Berechnung der EMD mithilfe der Ordered Distance. Dieser Algorithmus wurde aus dem Rechenbeispiel zur Ordered Distance aus Li et al. [LLV07, S. 6] abgeleitet.

Algorithmus 4.2 : EMD für quantitative Attributwerte

Eingabe : Häufigkeitsverteilungen $X_E = \{f_E(x_1), \ldots, f_E(x_m)\}$ und $X_V = \{f_V(x_1, \ldots, f_V(x_m)\}$ von quantitativen Attributwerten
Ausgabe : Abstand $EMD(X_E \| X_V)$ von X_E zu X_V;

```
1  i ← 1
2  cost ← 0
3  while i ≤ m do
4      read f_E(x_i) ∈ X_E and f_V(x_i) ∈ X_V        /* rel. Häufigkeiten */
5      diff_i ← |f_V(x_i) − f_E(x_i)|               /* Differenz der Häufigkeiten */
6      f_E(x_i) ← f_V(x_i)                          /* Transformation von f_E(x_i) */
7      if f_V(x_i) > f_E(x_i) then                  /* (NB2) */
8          f_E(x_{i+1}) ← f_E(x_{i+1}) − diff_i
9      else
10         f_E(x_{i+1}) ← f_E(x_{i+1}) + diff_i
11     end
12     cost ← cost + (1/(m−1) · diff_i)             /* c(x_i, x_{i+1}) f(x_i, x_{i+1}) */
13     i ← i + 1
14 end
15 return d_EMD(X_E ‖ X_V) ← cost
```

Grundlage des Algorithmus ist die von Li et al. formulierte Aussage, dass ein optimaler, durch die Ordered Distance berechneter Fluss, allein durch die sequenzielle Bewegung von Wahrscheinlichkeitsmasse zwischen adjazenten Elementen erfolgen kann [LLV07, S. 6]. Dies folgt aus der Tatsache, dass eine Bewegung von Wahrscheinlichkeitsmasse über mehr als zwei Elemente hinweg in Bewegungen zwischen benachbarten Elementen zerlegt werden kann [LLV07, S. 6]. Es stellt sich die Frage, wie mit Elementen zu verfahren ist, deren Nachbarn nicht die erforderliche Wahrscheinlichkeitsmasse aufweisen, um die benötigte Menge zu bedienen. Dies ist in folgendem Beispiel der Fall:

Beispiel 11. Seien E und V zwei Stichproben mit den Häufigkeitsverteilungen $X_E = \{0, 0, \ldots, 0, 1\}$ sowie $X_V = \{1, 0, 0, \ldots, 0\}$. Algorithmus 4.2 behandelt in diesem Fall die relativen Häufigkeiten nicht als Häufigkeiten im Sinne der deskriptiven Statistik,

denn diese sieht zum einen keine Arithmetik auf Häufigkeitsverteilungen vor und erlaubt zum anderen keine negativen Häufigkeitswerte. Das Verfahren von Li et al. fasst die Verteilungen daher als Menge von reellen Zahlen über einem metrischen Raum auf und erlaubt somit die Addition sowie Subtraktion. Die Behandlung des Elements an Position x_{i+1} hängt davon ab, ob das nachzubildende Element mehr Wahrscheinlichkeitsmasse erfordert als das aktuelle Element x_i besitzt. Diese Unterscheidung wird in Zeile 7 des Algorithmus getroffen. Besitzt es mehr, so bedeutet dies – in der Analogie der EMD ausgedrückt – das Erdloch, welches die Ausprägung an Position x_{i+1} darstellt und welches es zu füllen gilt, etwas tiefer auszuheben. Anschließend wird das Loch seinerseits von der Erdmasse seines Nachbarn aufgefüllt. Nach der ersten Iteration des Algorithmus, hätte die Mengenrepräsentation der Verteilung X_E die Form $\{1, -1, 0, \ldots, 1\}$. Nach der zweiten $\{1, 0, -1, \ldots, 1\}$ u. s. w. Letztendlich wird die Form $X_E = \{1, 0, 0, \ldots, 0\} = X_V$ erreicht.

Die sequentielle Umformung aller Elemente einer Stichprobe in die Elemente der anderen Stichprobe bildet einen minimalen Fluss [LLV07]. Die Kostenberechnung aus Algorithmus 4.2 lässt sich nach Li et al. wie folgt zusammenfassen:

$$\begin{aligned} d_{EMD}(X_E \| X_V) &= \frac{1}{m-1}(diff_1 + (diff_1 + diff_2) + \ldots + \\ &\qquad + (diff_1 + \ldots + diff_m)) \\ &= \frac{1}{m-1} \sum_{i=1}^{m} \sum_{j=1}^{i} diff_j \end{aligned} \tag{4.9}$$

Diese Berechnungsvorschrift erfüllt die Bedingungen der EMD für quantitative Attributwerte. Nachfolgend wollen wir uns der Berechnung der EMD für kategoriale Attributwerte zuwenden.

4.3.2.4 *Berechnung der EMD für kategoriale Attributwerte*

Die Notwendigkeit bei der Anonymisierung von sensiblen Attributwerten zwischen kategorialen und quantitativen Attributwerten zu unterscheiden, wurde ebenfalls von anderen Autoren erkannt [Zha+07]. Zhang et al. sahen den Schutz kategorialer Attributwerte durch das Konzept der ℓ-Diversity gegeben. Dass jedoch die Anonymität kategorialer Attributwerte durch Attribute Disclosure bedroht ist, wurde bereits in Abschnitt 4.2.1 gezeigt.

Li et al. unterscheiden zwei mögliche Abstandsmaße für kategoriale Attribute. Erstgenannt sei die *Equal Distance*. Für dieses Distanzmaß ist die Ground-Distance

zwischen allen Attributwerten einheitlich mit 1 definiert. Hierdurch ergibt sich nach Li et al. für den minimalen Massefluss folgende Berechnungsvorschrift:

$$d_{EMD}(X_E \| X_V) = \frac{1}{2}\sum_{i=1}^{m}|f_E(x_i) - f_V(x_i)| = \sum_{f_E(x_i)\geq f_V(x_i)}(f_E(x_i) - f_V(x_i)) \\ = -\sum_{f_E(x_i)<f_V(x_i)}(f_E(x_i) - f_V(x_i)) \tag{4.10}$$

Diese Gleichung ist als *totale Variation* bekannt. Sie vernachlässigt ebenfalls das Element der semantischen Nähe [LLV07, S. 5]. Li et al. entwickelten eine weitere Möglichkeit der Distanzberechnung, welche als *Hierarchical Distance* bekannt ist. Grundlage für dieses Abstandsmaß ist die Verwendung eines VGH (vgl. Kapite 3) über den sensiblen Attributwerten. Eine mögliche Darstellung eines VGH für das Attribut `Krankheit` von Tabelle 3.2 ist in Abbildung 4.3 gegeben.

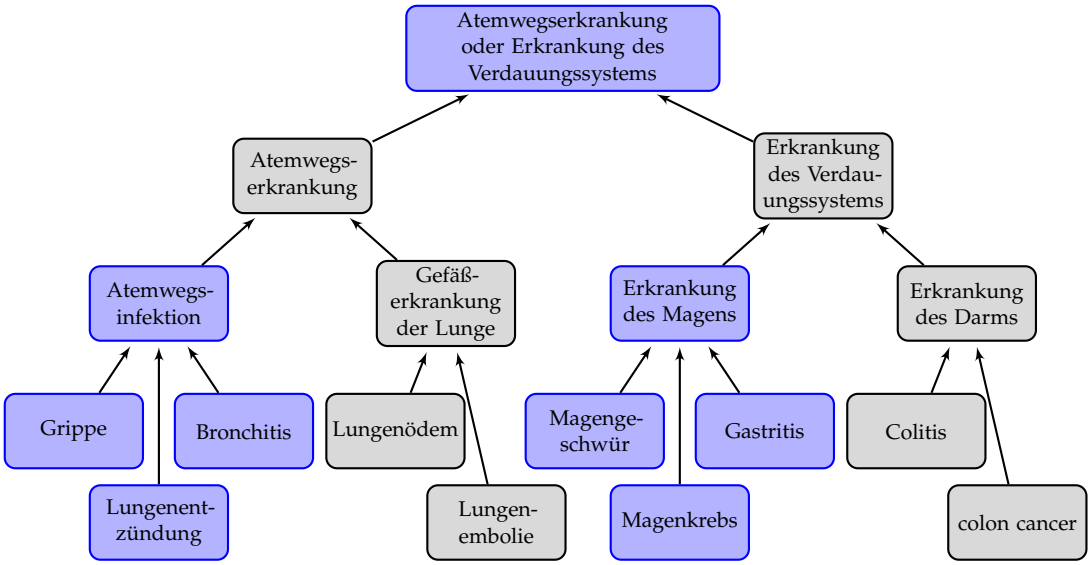


Abbildung 4.3: Generalisierungshierarchie des kategorialen Attributs `Krankheit`

Bevor wir uns der Berechnung eines optimalen Flusses zuwenden, werden wir die Motivation hinter der Verwendung eines VGH betrachten. Die Grundidee der Hierarchical Distance fußt auf der Annahme, dass sich Attributwerte – in der VGH als Blätter dargestellt – nahe sind, wenn sie einen *Lowest Common Ancestor* auf einer niedrigen Ebene in der VGH besitzen. Der Lowest Common Ancestor zweier Blätter x_i und x_j beschreibt denjenigen inneren Knoten eines gewurzelten Baums B, welcher

Vorgänger beider Blattknoten und am weitesten von der Wurzel $root(B)$ entfernt ist [Gus97, S. 181]. Wir werden diesen mit $lcs(x_i, x_j)$ notieren und bestimmen ferner die Höhe eines Knotens K im VGH mittels $height(K)$. Die Höhe eines Knotens in einem VGH beschreibt, im Kontext dieser Arbeit, die Länge eines Weges zu einem Blatt unterhalb des Teilbaums mit der Wurzel K. Diesen Teilbaum wollen wir im Folgenden mit $subtree(K)$ bezeichnen. Es ist in diesem Fall unerheblich welches Blatt als Maß herangezogen wird, da alle Blätter eine identische Entfernung zur Wurzel aufweisen. Folglich besitzt ein Blattknoten die Höhe 0 und dessen direkter Vorgänger die Höhe 1. Die Höhe des VGH sei durch $height(root(VGH))$ festgelegt. Die Ground-Distance zweier Attributwerte ergibt sich nach Li et al. aus:

$$c(x_i, x_j) = \frac{height(lcs(x_i, x_j))}{height(root(VGH))} \tag{4.11}$$

Auf Grundlage dieser Annahme werden wir im Folgenden Hilfsfunktionen definieren, welche es uns gestatten, einen Algorithmus zur Berechnung des optimalen Masseflusses aufzustellen. Seien erneut X_E und X_V zwei Häufigkeitsverteilungen eines sensiblen Attributs X. Zunächst bezeichne $child(K)$ die Menge aller Blätter des Teilbaums mit der Wurzel K. So ist beispielsweise $child(\texttt{Atemwegsinfektion}) = \{\texttt{Grippe, Lungenentzündung, Bronchitis}\}$. Des Weiteren definieren wir die Funktion

$$extra(K) = \begin{cases} f_E(x_i) - f_V(x_i) & K \text{ ist Blattknoten} \\ \sum_{c \in child(K)} extra(c) & \text{sonst} \end{cases} \tag{4.12}$$

Diese rekursive Funktion berechnet die Wahrscheinlichkeitsmasse, welche zum Erlangen des optimalen Flusses durch einen Knoten transportiert werden muss [LLV07, S. 7]. Für einen Blattknoten beträgt sie genau die Differenz zwischen der Ausgangsverteilung und der zu erreichenden Verteilung. Bei einem inneren Knoten beträgt sie die Summe der zu transportierenden Wahrscheinlichkeitsmasse der Blattknoten des Teilbaums, welcher den Knoten zur Wurzel hat.

Der Fluss der Wahrscheinlichkeitsmasse kann positiv oder negativ sein, je nachdem, ob überschüssige Masse vorhanden ist oder nicht. Für die Berechnung des optimalen Flusses ist es erforderlich, zwischen positiven und negativen Flüssen zu unterscheiden. Positive Flüsse verlassen einen inneren Knoten K in Richtung Wurzelknoten und negative Flüsse bezeichnen die Menge an Wahrscheinlichkeitsmasse, die der Teilbaum mit der Wurzel K erhalten muss. Die Unterscheidung zwischen

positiven und negativen Flüssen wird mithilfe der Funktion $extra()$ getroffen. Nachfolgend sind zwei Funktionen beschrieben, welche diesen Umstand abbilden:

$$pos\_extra(K) = \sum_{c \in \mathcal{C}} |extra(c)|, \text{ mit } \mathcal{C} = \{c \in child(K) : extra(c) > 0\} \tag{4.13}$$

$$pos\_extra(K) = \sum_{c \in \mathcal{C}} |extra(c)|, \text{ mit } \mathcal{C} = \{c \in child(K) : extra(c) < 0\} \tag{4.14}$$

Der Rückgabewert dieser Funktionen ist durch den Betrag stets positiv. Für die Interpretation dieser Werte müssen wir jeweils zwei Fälle unterscheiden:

1. $pos\_extra(K)$ ist größer als $neg\_extra(K)$
   a) $neg\_extra(K)$ ist größer als 0
   b) $neg\_extra(K)$ ist gleich 0
2. $pos\_extra(K)$ ist kleiner als $neg\_extra(K)$
   a) $neg\_extra(K)$ ist größer als 0
   b) $neg\_extra(K)$ ist gleich 0

Für den Fall, dass beide Werte von 0 verschieden sind (Punkt 1a), hat der Teilbaum mit der Wurzel K einen Überschuss an Wahrscheinlichkeitsmasse, d. h. es gilt die Ungleichung:

$$\sum_i f_E(x_i) > \sum_i f_V(x_i),\ x_i \in subtree(K) \tag{4.15}$$

Gleichzeitig gilt jedoch für *mindestens* ein $x_i \in subtree(K)$:

$$f_E(x_i) < f_V(x_i) \tag{4.16}$$

Dies ergibt sich direkt aus den Definitionen der Gleichungen (4.14) sowie (4.12).

Dieses Defizit innerhalb des Teilbaums kann gemäß Gleichung (4.15) von anderen Blättern des Teilbaums ausgeglichen werden. Der verbleibende Überschuss kommt bei der Kostenberechnung eines höher gelegenen Knotens zum Tragen.

Ist $neg\_extra = 0$ (Punkt 1b), so besteht ebenfalls ein Masseüberschuss in diesem Teilbaum, welcher bei der Kostenberechnung eines höher gelegenen Knotens berücksichtigt wird. Es erfolgt jedoch keine Umverteilung innerhalb $subtree(K)$, da Bedingung (4.16) nicht erfüllt ist.

In den verbleibenden zwei Fällen ist der Bedarf an Wahrscheinlichkeitsmasse des Teilbaums größer als seine Ressourcen. Im Fall des Punktes 2a gilt jedoch für ein Blatt x_i des Teilbaums: $f_E(x_i) > f_V(x_i)$. Dieses Blatt kann seine überschüssige Kapazität an andere Blätter des Teilbaums verteilen. Im Fall 2b finden ausschließlich

Massebewegungen in den Teilbaum hinein statt. Die hierbei entstehenden Kosten werden bei der Kostenberechnung eines höheren Knotens berücksichtigt.

Bisher wurden in diesem Zusammenhang noch keine Angaben zu den entstehenden Kosten für das Bewegen von Wahrscheinlichkeitsmasse zwischen den Blättern des VGH gemacht. In Gleichung (4.11) wurde bereits eine Kostenfunktion gegeben. Zusammen mit den Aussagen 1a bis 2b ergibt sich nach Li et al. folgende Kostenberechnung für einen inneren Knoten K:

$$cost(K) = \frac{height(K)}{height(root(VGH))} \cdot \min(pos\_extra(K), neg\_extra(K)) \tag{4.17}$$

Summieren wir die Kosten über alle inneren Knoten K des VGH, so erhalten wir nach Li et al. die EMD für kategoriale Attributwerte:

$$d_{EMD}(E\|V) = \sum_{K} \frac{height(K)}{height(root(VGH))} \cdot \min(pos\_extra(K), neg\_extra(K)) \tag{4.18}$$

Diese Gleichung unterschlägt den in den Punkten 1a und 2a erwähnten Ausgleich an Wahrscheinlichkeitsmasse. Diese ist für die korrekte Berechnung der Kosten jedoch unumgänglich. Algorithmus 4.3 bietet daher eine Umsetzung der Gleichung (4.18) unter Berücksichtigung der Veränderungen der Häufigkeitsverteilungen. Zeile 5 des Algorithmus bedient diesen Umstand, ohne sich in die Details der Verteilung zu vertiefen. Diese sind für die Kostenberechnung irrelevant, da gemäß Gleichung (4.11) alle Blattknoten eines Teilbaums denselben Abstand zueinander haben. Nachfolgend soll die beispielhafte Berechnung der t-Closeness für die Äquivalenzklasse E_2 aus Tabelle 4.2 die Anwendung des Algorithmus 4.3 verdeutlichen.

Beispiel 12. Der VGH sei wie in Abbildung 4.3. Die Häufigkeitsverteilung des sensiblen Attributs $X =$ `Krankheit` der Äquivalenzklasse E_2 betrage gemäß Tabelle 4.3 $X_{E_2} = \{1/3, 0, 1/3, 0, 0, 1/3\}$. Die Verteilung der gesamten Tabelle V sei $X_V = \{1/9, 1/9, 2/9, 1/9, 2/9, 2/9\}$. Wir erstellen zunächst eine Liste der inneren Knoten des VGH. Diese sei mit $\mathcal{I} =$ {`Atemwegserkrankung`, `Erkrankung des Magens`, `Gefäßerkrankung der Lunge`, `Erkrankung des Darms`, $root(VGH)$} bezeichnet.

Sei im ersten Schritt des Algorithmus $K =$ `Atemwegserkrankung`. Der Übersichtlichkeit halber werden wir zunächst die Funktion $extra()$ für die Blattknoten unterhalb von K sowie den Knoten K selbst berechnen:

$$\begin{aligned}
extra(\texttt{Grippe}) &= extra(x_1) = f_{E_2}(x_1) - f_V(x_1) = 1/3 - 1/9 = 2/9 \\
extra(\texttt{Lungenent.}) &= extra(x_2) = f_{E_2}(x_2) - f_V(x_2) = 0 - 1/9 = -1/9 \\
extra(\texttt{Bronchitis}) &= extra(x_3) = f_{E_2}(x_3) - f_V(x_3) = 1/3 - 2/9 = 1/9 \\
extra(K) &= \sum_{C \in child(K)} extra(C) = 2/9 + (-1/9) + 1/9 = 2/9
\end{aligned} \tag{4.19}$$

Algorithmus 4.3 : EMD für kategoriale Attributwerte

Eingabe : Häufigkeitsverteilungen $X_E = \{f_E(x_1), \ldots, f_E(x_m)\}$ und $X_V = \{f_V(x_1, \ldots, f_V(x_m)\}$ von kategorialen Attributwerten
Ausgabe : Abstand $EMD(X_E \| X_V)$ von X_E zu X_V;

```
1 cost ← 0
2 foreach innerer Knoten K do
3     min ← min(pos_extra(K), neg_extra(K))
4     if min > 0 then
5         verteile min auf Blätter von K
6         cost ← cost + height(K)/height(root(VGH)) · min
7     end
8 end
9 return d_EMD(X_E || X_V) ← cost
```

Ausgehend von diesen Berechnungen stellen wir mithilfe der Funktionen (4.13) und (4.14) fest, in welche Richtung Wahrscheinlichkeitsmasse transportiert werden muss.

$$\begin{aligned} pos\_extra(K) &= \sum_{c \in \mathcal{C}} |extra(c)| = |extra(x_1) + extra(x_3)| \\ &= 2/9 + 1/9 = 1/3 \\ neg\_extra(K) &= \sum_{c \in \mathcal{C}} |extra(c)| = |extra(x_2)| = 1/9 \end{aligned} \tag{4.20}$$

An dieser Stelle ist ersichtlich, dass die Verteilung X_{E_2} im Teilbaum $subtree(K)$ mehr Wahrscheinlichkeitsmasse besitzt als ihre Entsprechung X_V. Gleichzeitig hat das Element x_2 ein Defizit, welches auszugleichen ist. Wir befinden uns daher in der Situation, wie sie in Punkt 1a geschildert ist. Wir berechnen nun das Minimum der beiden Werte $pos\_extra()$ und $neg\_extra()$. Dadurch erhalten wir den Wert, welcher von einem Blatt des Teilbaums $subtree(K)$ zu einem andern Blatt „bewegt“ werden muss.

$$\min(pos\_extra(K), neg\_extra(K)) = \min\left(\frac{1}{3}, \frac{1}{9}\right) = \frac{1}{9} \tag{4.21}$$

O. B. d. A. subtrahieren wir $1/9$ von $f_{E_2}(x_3)$ und addieren diese zu $f_{E_2}(x_2)$. Es ergibt sich das folgende Bild der Verteilungen:

$$X_{E_2} = \{1/3, 1/9, 2/9, 0, 0, 1/3\} \tag{4.22}$$

$$X_V = \{1/9, 1/9, 2/9, 1/9, 2/9, 2/9\} \tag{4.23}$$

Als entstehende Kosten notieren wir:

$$cost = \frac{height(K)}{height(root(K))} \cdot min = \frac{1}{3} \cdot \frac{1}{9} = \frac{1}{27} \tag{4.24}$$

Wir werden an dieser Stelle die Kostenberechnung der restlichen Knoten überspringen und uns direkt der Wurzel zuwenden. Die Verteilung von E_2 hat unterdessen die Form $\{1/3, 1/9, 2/9, 1/9, 0, 2/9\}$ angenommen. Die Kosten der bisherigen Operationen betragen $cost = 2/27$. Die bekannten Funktionen liefern:

$$\begin{aligned} extra(x_1) &= 1/3 - 1/9 = 2/9 \\ extra(x_2) &= 1/9 - 1/9 = 0 \\ extra(x_3) &= 2/9 - 2/9 = 0 \\ extra(x_4) &= 1/9 - 1/9 = 0 \\ extra(x_5) &= 0 - 2/9 = -2/9 \\ extra(x_6) &= 2/9 - 2/9 = 0 \\ pos\_extra(root(VGH)) &= |extra(x_1)| = 2/9 \\ neg\_extra(root(VGH)) &= |extra(x_5)| = 2/9 \end{aligned} \tag{4.25}$$

Das Minimum $\min(pos\_extra(root(VGH)), neg\_extra(root(VGH)))$ beträgt $2/9$. Da beide Werte von 0 verschieden sind, wird als letztmöglicher Schritt $2/9$ von $f_{E_2}(x_1)$ subtrahiert und zu $f_{E_2}(x_5)$ addiert. Die Transformation ist abgeschlossen. Die entstandenen Kosten belaufen sich auf $2/27 + (3/3 \cdot 2/9) = 8/27$. Folglich beträgt die EMD zwischen E_2 und V:

$$d_{EMD}(E_2 \| V) = \frac{8}{27} \tag{4.26}$$

E_2 besitzt demnach t-Closeness für ein t von mindestens $8/27$. Wir sagen auch: „E_2 ist $8/27$-close zu V“.

Wir halten nach diesem Beispiel noch einmal das Folgende fest: Der Faktor t gibt das gewünschte maximale Ergebnis der Funktion $d_{EMD}(\cdot, \cdot)$ für jede Äquivalenzklasse an. Er nimmt Werte zwischen 0 und 1 an. Wobei der Wert 0 dem Vergleich zweier identischer Verteilungen entspricht und der Wert 1 zweier maximal verschiedener Verteilungen.

Im Kontext der t-Closeness-Berechnung einer Äquivalenzklasse zur Gesamtveröffentlichung wird der Wert von 1 nie erreicht. Dies liegt in der Tatsache begründet, dass die Verteilungen, welche durch die EMD hinsichtlich ihres Abstandes untersucht werden, derselben Grundgesamtheit entstammen. Wir werden dies nachfolgend zeigen.

Theorem 2 (Die t-Closeness einer Veröffentlichung ist stets kleiner als 1)**.** *Sei die Häufigkeitsverteilung der sensiblen Attributwerte einer Veröffentlichung V gegeben. Bezeichne E die Häufigkeitsverteilung einer Teilmenge der durch V repräsentierten Ausprägungen, d. h. E sei eine Äquivalenzklasse. Dann ist der maximale Abstand $d_{EMD}(E\|V) < 1$.*

Beweis. Da E eine Teilmenge Merkmalsausprägungen von V besitzt, kommt jede Ausprägung x_i mit einer relativen Häufigkeit von $f_E(x_i) > 0$ auch in V mit einer relativen Häufigkeit von $f_V(x_i) > 0$ vor.

Dementsprechend gilt für alle $x_i : |f_E(x_i) - f_V(x_i)| < 1$. Denn für $|f_E(x_i) - f_V(x_i)| = 1$ müsste gelten, dass eines der beiden Merkmale als sicheres Ereignis auftritt und das andere unmögliches Ereignis. Es gilt also $f_E(x_i) = 0$ und $f_V(x_i) = 1$ oder $f_V(x_i) = 1$ und $f_E(x_i) = 0$.

Da nun $f(x_i, x_j) < 1$ für alle x_i, x_j, folgt aus Gleichung 4.5 und Nebenbedingung NB3, dass $d_{EMD}(E\|V) < 1$. □

4.3.3 *Schwächen der t-Closeness*

Das Konzept der t-Closeness wird vielfach hinsichtlich der Interpretation seines Parameters t kritisiert [LLV07; Cao+11; FZ08]. Kern der Kritik ist der nicht ersichtliche Zusammenhang zwischen t und dem erreichten Schutz der Privatsphäre. Das nachfolgende Beispiel aus [LLV07] erläutert diesen Punkt:

Beispiel 13. Seien A und B zwei Stichproben des Merkmals X vom Umfang $n = 100$, mit $m = 2$ Ausprägungen. Die Ausprägungen seien mit $X_A = \{f_A(x_1) = {}^1/_{100}, f_A(x_2) = {}^{99}/_{100}\}$ sowie $X_B = \{f_B(x_1) = {}^{11}/_{100}, f_B(x_2) = {}^{89}/_{100}\}$ gegeben. Berechnen wir die t-Closeness zwischen den Verteilungen mithilfe der Ordered Distance, so ergibt sich nach Gleichung 4.9 folgendes Ergebnis:

$$\begin{aligned} d_{EMD}(A\|B) &= \frac{1}{m-1}\sum_{i=1}^{m}\left|\sum_{j=1}^{i} f_A(x_j) - f_B(x_j)\right| \\ &= \frac{1}{2-1}\sum_{i=1}^{2}\left|\sum_{j=1}^{i} f_A(x_j) - f_B(x_j)\right| \\ &= 1 \cdot \left(\left|\frac{1}{100} - \frac{11}{100}\right| + \left|-\frac{10}{100} + \frac{99}{100} - \frac{89}{100}\right|\right) \\ &= 1 \cdot \frac{1}{10} + 0 = 0,1 \quad . \end{aligned}$$

Seien nun A' und B' zwei weitere Stichproben des Merkmals X mit ebenfalls $m = 2$ Ausprägungen. Die Häufigkeitsverteilungen seien wie folgt: $X_{A'} = \{f_{A'}(x_1) =$

$^4/_{10}, f_{A'}(x_2) = {}^6/_{10}\}$ sowie $X_{B'} = \{f_{B'}(x_1) = {}^5/_{10}, f_{B'}(x_2) = {}^5/_{10}\}$. Es fällt auf, dass die Ordered Distance ebenfalls einen Wert von $0,1$ liefert, denn

$$\begin{aligned} d_{EMD}(A' \| B') &= \frac{1}{m-1} \sum_{i=1}^{m} \left| \sum_{j=1}^{i} f_{A'}(x_j) - f_{B'}(x_j) \right| \\ &= 1 \cdot \left(\left| \frac{4}{10} - \frac{5}{10} \right| + \left| -\frac{1}{10} + \frac{6}{10} - \frac{5}{10} \right| \right) \\ &= \frac{1}{10} = 0,1 \quad . \end{aligned}$$

Somit sind beide Verteilungen $0,1$-close. Dies widerspricht dem intuitiven Verständnis der Nähe dieser Verteilungen. Der Informationsgewinn durch die Transformation von $0,01$ in $0,11$ entspricht einer um $1000\,\%$ erhöhten Wahrscheinlichkeit die Ausprägung x_1 zu besitzen. Wohingegen die Transformation von $0,4$ nach $0,5$ nur einem Zuwachs von $25\,\%$ entspricht. Die Verwendung der KL-Distanz als Kostenfunktion würde diesen Nachteil laut Li et al. nicht aufweisen, böte jedoch nicht die Vorteile, welche durch die Betrachtung der semantischen Nähe entstehen [LLV07, S. 10].

Das Konzept der t-Closeness ist eine Erweiterung der k-Anonymisierung, ebenso wie das der ℓ-Diversity [TG12, S. 11:2]. Ohne eine vorherige k-Anonymisierung können die Konzepte nicht angewandt werden. Dies folgt aus der Tatsache, dass es eines k-Anonymisierung-Algorithmus bedarf, um Tupel-Äquivalenzklassen zu erzeugen. Vielfach setzen Algorithmen für die erweiterten Konzepte ℓ-Diversity und t-Closeness auf etablierten Algorithmen zur Berechnung der k-Anonymisierung. Der Fokus dieser Algorithmen liegt dabei auf der Korrektheit des Verfahrens. Optimalität im Hinblick auf die Utility der Daten wird im Allgemeinen nicht verlangt. Hinsichtlich der k-Anonymisierung weit verbreitet ist der `Mondrian`-Algorithmus [LDR06]. Machanavajjhala et al. stellten in ihrer Veröffentlichung zur ℓ-Diversity den ℓ-Diversity-`Incognito`-Algorithmus vor [Mac+07]. Bezüglich der t-Closeness existiert eine Erweiterung des `Incognito`-Algorithmus [LLV10]. Es existieren Algorithmen, welche die gewünschte Eigenschaft des jeweiligen Konzepts zusammen mit der k-Anonymisierung herstellen. Beispielhaft seien `Decomposition` für ℓ-Diversity [Ye+09] sowie `SABRE` für t-Closeness [Cao+11] genannt.

Diverse Autoren griffen das Problem der Nicht-Optimalität der Resultate auf. Auch Li et al. notierten in der ursprünglichen Veröffentlichung t-Closeness die starke Einschränkung der Utility der Daten. So werden die Homogeneity Attack sowie die Background Knowledge Attack auf die k-Anonymisierung nicht verhindert, sondern vielmehr erschwert [LLV07, S. 7]. Diese Feststellung soll den Wert der t-Closeness nicht schmälern. Eine Folgerung der t-Closeness ist es, dass diese Angriffe ebenfalls auf eine maximal anonymisierte Tabelle möglich wären [LLV07, S. 7].

Es bleibt jedoch kritisch anzumerken, dass die Erweiterung der *k*-Anonymisierung um eine *t*-Closeness-Eigenschaft der Utility der Daten nicht zuträglich ist [FZ08]. Die Utility ist in vielen Fällen durch das eigentliche Verfahren bereits stark eingeschränkt [GCG08; BS08].

5 DEZENTRALE ANONYMISIERUNGSKONZEPTE

In den vorhergehenden Kapiteln wurden die Grundlagen für die Anwendung von Datenanonymisierungskonzepten des PPDP in einer verteilten Umgebung gelegt. Unser Hauptaugenmerk gilt der Überführung des Konzepts der t-Closeness in ein verteiltes Szenario. Wir konzentrieren uns auf vertikal partitionierte Daten.

Aus dem vorherigen Kapitel ist uns bereits bekannt, dass es einer k-Anonymisierung bedarf, um das Konzept der t-Closeness anzuwenden. Wir werden daher zunächst ein Konzept zur verteilten Datenanonymisierung unter Verwendung der k-Anonymisierung betrachten. Darauf aufbauend werden wir konzeptionell einen Algorithmus entwickeln, welcher es uns ermöglicht, eine gemeinsame Veröffentlichung der Daten unter Einhaltung der t-Closeness zu erreichen.

Methoden zur Berechnung einer ℓ-Diversity über verteilten Daten werden wir in Kapitel 6 betrachten.

5.1 TERMINOLOGIE UND ANNAHMEN ZUR VERTEILTEN DATENANONYMISIERUNG

Die Anonymisierung räumlich getrennter Datensätze verlangt die Einführung einer speziellen Terminologie, welcher wir uns nachfolgend widmen wollen.

5.1.1 *Terminologie*

In Kapitel 1 wurde der Anwendungsfall eines verteilten Szenarios mit der Existenz von multizentrischen Studien motiviert. Wir wollen diesen Anwendungsfall nun konkretisieren.

Im Folgenden gehen wir von zwei beteiligten medizinischen Institutionen aus. Wir werden diese als *Parteien* referenzieren und bezeichnen Partei 1 mit mit P_1 und Partei 2 mit P_2. Das Ziel beider Parteien ist die Veröffentlichung einer gemeinsamen Tabelle, welche einer vorgegebenen Anonymisierung genügt. Die Daten seien im Vorfelde der Anonymisierung vertikal partitioniert. Wir wollen den Begriff der vertikalen Partitionierung an dieser Stelle konkretisieren und definieren:

Definition 29 (Vertikale Partitionierung von Mikrodaten)**.** *Gegeben eine Tabelle T mit Attributen $A_1, \ldots, A_n$. Diese seien in den Mengen ID_T, QI_T sowie SA_T organisiert. Dann bezeichnet eine* vertikale Partitionierung *der Tabelle T die Zerlegung von T bezüglich der*

Attribute QI_T und SA_T in disjunkte Teilschemata $T^i(ID_T, QI_i, SA_i), 1 \leq i \leq m < n$, mit $|QI_i| \geq 1, |SA_i| \geq 1$ so dass gilt:

$$T = T^1 \bowtie_{ID_T} \ldots \bowtie_{ID_T} T^m \tag{5.1}$$

Der Einfachheit halber wollen wir annehmen, dass die ID eines Tupels parteiübergreifend eindeutig sei. Das bedeutet, dass durch die ID stets das gleiche Individuum referenziert wird. In den Abbildungen wird die ID grau dargestellt. Dies soll uns verdeutlichen, dass sie nicht Teil der Veröffentlichung ist. Die ID wird ausschließlich zu Zwecken der Nachvollziehbarkeit in den Abbildungen verwendet.

Dementsprechend verfügen beide Parteien über denselben Patientenstamm, jedoch unterschiedliche Attribute. Dabei bezeichne $T^i, i \in \{1, 2\}$ die Daten der jeweiligen Partei i. Des Weiteren gelten die Notationen aus Kapitel 3.

Da die Daten zwei getrennt erhobener Datensätze zusammengeführt werden, ist es zweckdienlich zwischen *getrennten* und *gemeinsamen* Daten zu unterscheiden. Wir wollen in Übereinstimmung mit der Literatur zur verteilten Datenanonymisierung nachfolgend von einem *lokalen* Attribut sprechen, wenn dieses ausschließlich in dem Schema einer Partei vorhanden ist. Gleichsam beschreibe ein *globales* Attribut ein Attribut, welches in den Schemata beider Parteien enthalten ist [Fun+11, S. 6].

Die Verteilung der Daten und die Berechnung einer Anonymisierung ohne die Preisgabe sensibler Informationen verlangt die Existenz kryptographischer Methoden. Wenngleich sich diese Arbeit auf die konzeptionellen Belange der verteilten Datenanonymisierung konzentriert, so ist ein Exkurs in dieses Fachgebiet unumgänglich. Wir werden dieses jedoch nur so weit erläutern, wie es zum Verständnis der nachfolgend vorgestellten Konzepte vonnöten ist.

5.1.2 *Kryptographische Grundlagen*

Die angesprochene Verteilung der Daten verlangt eine *Kommunikation* zwischen den beteiligten Parteien. Diese Kommunikation erfolgt über einen potentiell unsicheren Kanal. Die Absicherung der über diesen Kanal gesendeten Nachrichten ist Gegenstand der *Kryptographie* [Scho6b, S. 1].

Eine unveränderte Nachricht wird als *Klartext* bezeichnet. Wird diese mit Mitteln der Kryptographie verändert, um ihren Inhalt zu verbergen, so sprechen wir von einem *Chiffretext*. Der Vorgang der Veränderung eines Klartext in einem Chiffretext wird auch als *Verschlüsselung* bezeichnet. Der umgekehrte Weg – die Wandlung eines Chiffretexts in einem Klartext – wird als *Entschlüsselung* oder *Dechiffrierung* bezeichnet.

Die Absicherung einer Nachricht kann viele Formen annehmen. Laut Schneier sind die wesentlichen Ansprüche der Kryptographie die *Geheimhaltung*, *Authentifizierung*, *Integrität* sowie *Verbindlichkeit* einer Nachricht [Scho6b, S. 2].

Wir werden uns im Folgenden ausschließlich mit der Geheimhaltung von Nachrichten beschäftigen. Dies bedeutet, dass nachfolgend davon ausgegangen wird, dass der Kommunikationskanal zwischen den Parteien sicher gegenüber äußerer Veränderung ist. Weiterhin wollen wir festlegen, dass es zu keinem Nachrichtenverlust kommen kann. Eine Nachricht wird daher stets unverändert übermittelt. Wir stellen jedoch keine Anforderungen an den Kanal bezüglich seiner Abhörsicherheit.

Wir haben im Vorfelde von Ver- und Entschlüsselung gesprochen, ohne zu erläutern wie dieser Vorgang vollzogen wird. Dies wollen wir nun nachholen.

Die Verschlüsselung eine Klartexts M sowie die Entschlüsselung eines Chiffretexts C erfolgen mithilfe *kryptographischer Algorithmen*. Nach Schneier ist ein kryptographischer Algorithmus eine mathematische Funktion. Diese erwartet zwei Parameter. Zum einen den zu ver-/entschlüsselnden Klar- oder Chiffretext, zum anderen einen *Schlüssel*. Wir werden den Schlüssel nachfolgend mit K bezeichnen. Sollte eine Unterscheidung der Schlüssel der jeweiligen Parteien notwendig werden, so bezeichne K_1 den Schlüssel von Partei 1 und K_2 den Schlüssel von Partei 2. Die Verwendung eines Schlüssels stellt die Generalität der Funktion sicher. Durch die Verwendung unterschiedlicher und nur der jeweiligen Partei bekannter Schlüssel, ist es den beteiligten Parteien erlaubt denselben Algorithmus zu verwenden, ohne dadurch die Sicherheit des Chiffretext zu gefährden [Scho6b, S. 3]. Häufig werden weitere Anforderungen an den verwendeten Schlüssel gestellt. Wir wollen an dieser Stelle jedoch nicht weiter auf die Kriterien zur Auswahl von Schlüsseln und deren Erzeugung eingehen. Wir nehmen stattdessen an, dass beide Parteien einen im Sinne des Verfahrens kryptografisch sicheren Schlüssel besitzen.

Im Allgemeinen werden für die Ver- und Entschlüsselung zwei verschiedene Funktionen benutzt [Scho6b, S. 2]. Für den Vorgang der Verschlüsselung definieren wir eine Verschlüsselungsfunktion $E : K \times M \to C$. Dementsprechend sei die Entschlüsselungsfunktion durch $D : K \times C \to M$ gegeben[1].

Um eine kompaktere Darstellung zu erhalten, definieren wir eine äquivalente Notation der Funktionen E und D wie folgt:

Notation (Funktionen zur Ver- und Entschlüsselung einer Nachricht)**.**

$$E(K, M) \equiv E_K(M) = C \tag{5.2}$$
$$D(K, C) \equiv D_K(C) = M \quad . \tag{5.3}$$

1 Wir unterschlagen an dieser Stelle Angaben über den Wertebereich von K, C und M sowie der jeweils verwendeten Funktion E und D. Wir wollen annehmen, dass diese „korrekt gewählt seien“.

Der Vollständigkeit halber muss zwischen *symmetrischen* und *asymmetrischen* Algorithmen unterschieden werden. Bei symmetrischen Algorithmen ist der Schlüssel zum Ver- und Entschlüsseln einer Nachricht meist identisch [Scho6b, S. 4]. Dies ist bei asymmetrischen Algorithmen nicht der Fall. Bei dieser Art von Algorithmus unterscheiden sich die Schlüssel zur Ver- und Entschlüsselung stets [Scho6b, S. 5]. Für den Kontext dieser Arbeit ist die Art des Algorithmus irrelevant. Wir verzichten daher auf eine Festlegung und eingehende Betrachtung.

Von besonderer Wichtigkeit ist hingegen eine Klasse von Algorithmen zur Ver- und Entschlüsselung von Nachrichten, welche als *kommutative Verschlüsselungsfunktionen* bekannt sind.

5.1.2.1 *Kommutative Verschlüsselung*

Nach Agrawal et al. [AES03] ist eine kommutative Verschlüsselungsfunktion eine Funktion mit den folgenden Eigenschaften:

Definition 30 (Kommutative Verschlüsselung (informal))**.** *Sei $E_K(M)$ eine Verschlüsselungsfunktion und K ein beliebiger Schlüssel. Dann ist $E_K(M)$ kommutativ genau dann, wenn gilt:*

1. *Für alle Schlüssel K, K' gilt: $E_K\left(E_{K'}(M)\right) = E_{K'}\left(E_K(M)\right)$*
2. *Die Funktion $E_K(M)$ ist bijektiv*
3. *Gegeben ein Schlüssel K, dann gilt: $E_K^{-1}(\cdot)$ ist in polynomialer Zeit berechenbar*
4. *Sei das Ergebnis $E_K(M) = C$ einer zufällig gewählten Nachricht M sowie die Nachricht M bekannt. Sei ferner $E_K(M') = C'$ bekannt, dann gilt: Aus C, C', M lässt sich M' nicht ableiten*

Betrachten wir die Punkte genauer, ohne auf formale Details eingehen zu wollen[2]. Punkt 1 verlangt, dass das Resultat einer zweifachen Anwendung der Verschlüsselungsfunktion $E_K(\cdot)$ mit verschiedenen Schlüsseln K, K' unabhängig von der Reihenfolge der Verschlüsselung ist. Dieser Punkt ist wesentlich für die Namensgebung der Verschlüsselungsfunktion, assoziiert er sie doch mit Kommutativität der Addition und Multiplikation mathematischer Körper (vgl. [Foro8, S. 12ff]). Punkt 2 fordert die Eindeutigkeit der Verschlüsselung. Durch die Forderung einer Bijektion besitzen je zwei verschiedene Klartexte M, M' nie identische Chiffretexte C, C'. Der dritte Punkt fordert das Vorhandensein einer Umkehrfunktion zur Verschlüsselung – die

2 Diese finden sich in Agrawal et al. [AES03, S. 3f]

Entschlüsselung. Diese muss mithilfe des Schlüssels in polynomialer Zeit durchführbar sein. Es darf sich bei der Verschlüsselungsfunktion also nicht um eine *Einwegfunktion* handeln. Der letztgenannte Punkt fordert, dass es aufgrund der Kenntnis eines Klartextes M und des dazugehörigen Chiffretextes C nicht möglich sein darf, für einen weiteren Chiffretext C' den entsprechenden Klartext M' in polynomialer Zeit zu berechnen.

Im diesem Kapitel werden wir regen Gebrauch von kommutativen Verschlüsselungsfunktionen machen. Der Übersichtlichkeit halber werden wir in den auftretenden Beispielen eine sehr einfache Verschlüsselungsfunktion wählen: Die Verschlüsselung mittels des XOR-Operators [Scho6b, S. 15ff]. Wir erhalten die nachfolgenden Funktionen:

$$E_K(M) = M \oplus K = C \tag{5.4}$$
$$D_K(C) = C \oplus K = M \quad . \tag{5.5}$$

Die von uns gewählte XOR-Verschlüsselung genügt den Punkten 1-3 der Definition 30. Sie ist laut Schneier als nicht sicher anzusehen [Scho6b, S. 15ff]. Ferner genügt sie nicht Punkt 4 aus Definition 30. Dies ist begründet mit der Tatsache, dass die XOR-Verschlüsselung anfällig gegenüber *Known-Plaintext*-Angriffen ist [Scho6b, S. 6]. Dies bedeutet, dass durch die Kenntnis des Chiffretextes sowie des Klartextes der Schlüssel wie folgt berechnet werden kann: $M \oplus C = K$. Aufgrund der bekannten Angriffe auf die XOR-Verschlüsselung ist diese formal nicht als kommutative Verschlüsselung nach Agrawal et al. [AES03] zu sehen. Wir wollen über dieses Manko zugunsten der Nachvollziehbarkeit der Beispiele hinwegsehen. Es bleibt festzuhalten, dass eine Umsetzung der nachfolgend vorgestellten Algorithmen unter realen Bedingungen eine sichere kommutative Verschlüsselungsfunktion bemühen sollte, wie sie in Agrawal et al. [AES03] und Pohlig et al. [PH78] zu finden ist.

Kommutative Verschlüsselungsfunktionen finden u. a. in *kryptografischen Protokollen* Verwendung. Nach Schneier dient ein kryptografisches Protokoll „der Durchführung einer bestimmten Aufgabe und besteht aus einer Folge von Aktionen an denen zwei oder mehr Parteien beteiligt sind“ [Scho6b, S. 25f]. Im Zuge der Protokoll-Ausführung liefert jede Partei eine Eingabe und erhält ein Ergebnis im Anschluss an die Durchführung des Protokolls.

In Abgrenzung zum Terminus „Algorithmus“, verwenden kryptographische Protokolle kryptographische Algorithmen. Dies geschieht zum Schutz der Kommunikation zwischen den Parteien.

Als Ideal eines kryptographischen Protokolls wird die Trusted-Third-Party (TTP) angesehen [Golo4, S. 601]. In diesem Modell kommunizieren sämtliche beteiligten Parteien mit einer vertrauenswürdigen dritten Partei. Dieser werden die Eingaben je-

der Partei übermittelt. Im Anschluss errechnet die Partei das Ergebnis des Protokolls und übergibt dieses den Parteien.

Wie in Jiang et al. [JC05, S. 167] dargelegt wurde, ist die Verwendung einer TTP nicht immer möglich oder sinnvoll. In diesem Fall werden Protokolle der *Secure-Multiparty-Computation (SMC)* verwendet, um eine TTP zu simulieren [Gol04, S. 601]. Ziel dieser Klasse von kryptographischen Protokollen ist die Berechnung einer Funktion in einem Mehrparteien-Szenario. Jede Partei liefert eine zur Berechnung der Funktion benötigte Eingabe. SMC-Protokolle nutzen kryptografische Funktionen, um die Eingabe jeder Partei vor den anderen Parteien zu verbergen. Daraus resultiert ihre fundamentale Forderung, den beteiligten Parteien durch die Ausführung des Protokolls keinen zusätzlichen Informationsgewinn zu ermöglichen. Ausgenommen ist die Information, welche aus der eigenen Eingabe und dem Ergebnis des Protokolls gefolgert werden kann [Gol04, S. 601f]. SMC-Protokolle stellen eine Verallgemeinerung des Secure-Two-Party-Problems dar [Gol04, S. 601f]. Dieses wurde in der Absicht formuliert, ein Verfahren zur Bestimmung des Maximalwertes der Eingaben zweier an dem Verfahren beteiligter Parteien zu entwickeln. Es wurde erstmalig von Yao [Yao82] verfasst und später als *¸Millionaire Problem* bekannt.

Ein in dem Zusammenhang mit SMC-Protokollen häufig erwähnter Begriff ist der Begriff der *Homomorphen Verschlüsselung*. Klartexte, welche mittels eines homomorphen Verschlüsselungsalgorithmus chiffriert wurden, erlauben algebraische Operationen auf den verschlüsselten Daten, ohne diese vorher entschlüsseln zu müssen [Gen09]. Sie eignen sich daher für die Verwendung in SMC-Protokollen. Die homomorphe Verschlüsselung ist für die Betrachtung der in dieser Diplomarbeit vorgestellten Protokolle nicht wesentlich. Wir werden uns daher nicht eingehender mit dieser beschäftigen.

Wir haben die Grundlagen für das Verständnis der verwendeten Kryptographie gelegt. Nachfolgend widmen wir uns Protokollen, welche die Berechnung einer Anonymisierung von vertikal partitionierten Daten erlauben. Diese Protokolle verwenden die kommutative Verschlüsselung zum Schutz der Daten.

5.2 k-ANONYMISIERUNG ÜBER VERTIKAL PARTITIONIERTEN DATEN

In diesem Abschnitt werden wir uns dem Protokoll „Distributed Privacy-Preserving two-Party Generic Anonymizer“ (DPP_2GA) von Jiang et al. im Detail widmen. Dieses Protokoll war das erste Protokoll, welches eine k-Anonymisierung von vertikal partitionierten Daten ermöglichte [MFD11, S. 586f].

5.2.1 DPP_2GA

2005 entwarfen Jiang et al. das Protokoll „DPP_2GA“ zur Berechnung einer k-Anonymisierung über vertikal partitionierten Daten. Dieses Protokoll verwendet Prinzipien der SMC, um dieses Ziel zu erreichen. Demgemäß definieren Jiang et al. Annahmen, nach welchen Gesichtspunkten das von ihnen entwickelte Protokoll als sicher anzusehen ist.

5.2.1.1 *Annahmen zur Ausführung des Protokolls*

Die zentrale Forderung betrifft den Schutz der Privatsphäre zu jedem Zeitpunkt der Ausführung des Protokolls. [S. 167 JC05, Definition 1]. Diese Anforderung erlaubt den an dem Protokoll beteiligten Parteien Einblicke in Zwischenergebnisse des Protokolls. Diese Einschränkung ermöglicht zum einen das Überprüfen der verteilten Daten auf eine valide k-Anonymisierung. Zum anderen stellt es einen Bruch mit den Prinzipien der SMC dar. Diese gestatten den Parteien keinen Erkenntnisgewinn durch die Protokollausführung [Gol04, S. 607ff]. Streng genommen handelt es sich bei dem DPP_2GA-Protokoll daher nicht um ein SMC-Protokoll. Jiang et al. zeigten jedoch, dass trotz dieser Einschränkung der Schutz der anonymisierten Daten zu jeder Zeit gewährleistet ist [JC05, S. 174f].

Die Sicherheit eines kryptografischen Protokolls hängt ebenfalls von der Modellierung eines *Angreifers* ab [Gol04, S. 603]. Im Kontext der SMC beschreibt ein Angreifer eine an der Ausführung des Protokolls beteiligte Partei[3] [Gol04, S. 602]. Die Modellierung eines Angreifers beschreibt somit das Verhalten der an dem Protokoll beteiligten Parteien. Diese Modellierung ist maßgeblich für die Sicherheit des Protokolls. Eine häufig getroffene Annahme ist die eines Angreifers nach dem *Honest-But-Curious*-Prinzip (HBC)[4]. Der Angreifer ist den beteiligten Parteien nicht bekannt. Im Gegensatz zu dem als *böswilligen* Angreifer bezeichneten Modell, folgt der Semi-Honest-Angreifer dem Ablauf des Protokolls. Weiterhin manipuliert ein derartiger Angreifer weder seine Eingabe noch etwaige Zwischenergebnisse. Es ist ihm jedoch gestattet die Zwischenergebnisse des Protokolls aufzuzeichnen und aus diesen Schlüsse zu ziehen [Gol04, S. 603]. Ein SMC-Protokoll darf demnach korrekte Eingaben erwarten, sofern es gegen ein HBC-Angreifermodell entwickelt wurde. Es muss jedoch sicherstellen, dass kein Zwischenergebnis Rückschlüsse auf die Eingaben der beteiligten Parteien zulässt.

3 Streng genommen, besteht keine Einschränkung bezüglich der Anzahl der Angreifer. Da wir jedoch ein Zwei-Parteien-Protokoll betrachten, ist es sinnvoll von einem Angreifer auszugehen.

4 Zuweilen wird es auch als *Semi-Honest*-Prinzip bezeichnet.

Auch das Protokoll von Jiang et al. geht von einem Angreifer nach dem HBC-Prinzip aus. Für das Anwendungsszenario dieser Diplomarbeit ist dies eine sinnvolle Annahme. Wir erinnern uns an das gemeinsame Ziel zweier medizinischer Institute, eine Veröffentlichung ihrer Ergebnisse vorzunehmen. Eine absichtliche Manipulation des Protokolls würde das gemeinsame Ziel und somit die Studie gefährden.

Das DPP_2GA-Protokoll ermöglicht die Berechnung einer gemeinsamen k-Anonymisierung zwischen zwei Parteien. Demgemäß verfügen beide Parteien über einen vertikal partitionierten Teil der Tabelle. Das Schema der Partei 1 sei gemäß Definition 29: $T^1(ID_T, QI_1, SA_1)$, für die lokalen Attribute QI_1 und SA_1 von Partei 1. Entsprechend sei das Schema von Partei 2: $T^2(ID_T, QI_2, SA_2)$.

Die umfangreiche Vorarbeit ermöglicht uns nun die Analyse des DPP_2GA-Protokolls.

5.2.1.2 *Berechnung einer gemeinsamen k-Anonymisierung*

Bislang ist uns zum Erstellen einer k-Anonymisierung ausschließlich der DF-Algorithmus von Sweeney bekannt (vgl. Kapitel 4). Dieser verwendet eine Heuristik zur Full-Domain-Generalisierung (FDG), um eine k-Anonymisierung herzustellen. Hierzu benötigt der Algorithmus Zugriff auf die QIDs der Mikrodaten.

Jiang et al. konnten zeigen, dass jede Teilmenge eines k-anonymen QID ebenfalls eine valide k-Anonymisierung besitzt [S. 170 JC05, Theorem 1]. Diese Erkenntnis ermöglicht das Berechnen einer k-Anonymisierung über den lokalen Daten. Wir wollen diese nachfolgend als lokale k-Anonymisierung referenzieren. Zum Erreichen einer k-Anonymisierung über den globalen Daten (globale k-Anonymisierung), gilt es zu überprüfen, ob die Vereinigung der lokalen Anonymisierungen erneut k-anonym ist. Anschließend können diese über der ID vereinigt werden. Dieses Prinzip der Anonymisierung von verteilten Daten ist als *Generalize-Then-Mashup* bekannt [Fun+11, S. 3].

Wie im vorherigen Abschnitt definiert, ist die zentrale Forderung des Protokolls der Schutz der Privatsphäre. Dies muss für jeden Ausführungszeitpunkt des Protokolls gelten. Würden Partei 1 und 2 jeweils eine lokale k-Anonymisierung erstellen, so wäre es ihnen unmöglich diese auszutauschen und anschließend auf Einhaltung einer globalen k-Anonymisierung zu überprüfen. Es bestünde die Möglichkeit, dass durch die lokale k-Anonymisierung Äquivalenzklassen über verschieden Tupeln erzeugt wurden. Bei einer Vereinigung der lokalen Anonymisierungen würden die Äquivalenzklassen den Umfang der Schnittmenge der enthaltenen Tupel aufweisen. Dieser könnte geringer als k sein. In der Folge wäre der Schutz der Mikrodaten durch die k-Anonymisierung nicht mehr gewährleistet.

Es gilt ein Verfahren zu entwickelt, mit dessen Hilfe ein Vergleich der Äquivalenzklassen stattfinden kann, ohne die Privatsphäre der durch die Mikrodaten repräsentierten Personen zu gefährden. Mithilfe der kommutativen Verschlüsselung ist es uns möglich dies zu realisieren.

Zu diesem Zweck definierten Jiang et al. ein Mengensystem γ^i. Dieses enthält Mengen von IDs der lokalen k-Anonymisierung mit den folgenden Eigenschaften:

Definition 31 (Eigenschaften des Mengensystems γ^i)**.** *Sei $\gamma^i[p]$ die $p-te$ Teilmenge von γ^i. Dann besitzen alle Einträge, deren IDs in $\gamma^i[p]$ enthalten sind, dieselben Werte in den QIDs. Zusätzlich gilt für jedes γ^i:*

- $\gamma^i[p] \cap \gamma^i[q] = \emptyset$ *für jedes* $1 \leq p, q \leq |\gamma^i|$ *mit* $p \neq q$
- $\bigcup_p \gamma^i[p] = \gamma^i$

Wir verfügen nun über eine Datenstruktur, welche uns eine Referenz auf die Äquivalenzklassen liefert. Anhand dieser ist es uns möglich, den Begriff des *Vergleichs zweier Mengensysteme* näher zu erläutern. Er wird für den nachfolgenden Algorithmus zentrale Bedeutung besitzen.

Definition 32 (Äquivalenz zweier Mengensysteme)**.** *Seien γ^i, γ^j zwei Mengensysteme. Bezeichne ferner $\gamma^i[p]$ die $p-te$ Teilmenge von γ^i und $\gamma^j[q]$ bezeichne die $q-te$ Teilmenge von γ^j. Zwei Mengensysteme γ^i, γ^j sind äquivalent genau dann, wenn für keinen nichtleeren Schnitt der in ihnen enthaltenen Mengen gilt:*

$$0 < |\gamma^i[p] \cap \gamma^j[q]| < k,\ \forall \gamma^i[p] \in \gamma^i, \gamma^j[q] \in \gamma^j \quad .$$

Wir schreiben $\gamma^i \equiv \gamma^j$.

Der Austausch und Vergleich der Mengensysteme unter Verwendung einer kommutativen Verschlüsselung ist in Algorithmus 5.1 dargestellt. Wir wollen diesen anhand eines Beispiels erläutern.

Beispiel 14. Im Folgenden betrachten wir ein Zwei-Parteien-Szenario, wie es in Tabelle 5.1 beschrieben ist. Die Extension von Partei 1 (P_1) sei in Abbildung 5.1a dargestellt. P_1 verfüge demnach über die QIDs „`PLZ, Alter`“ sowie das SA „`Krankheit`“. Die Extension von Partei 2 (P_2) entnehmen wir Abbildung 5.1b. Partei 2 besitzt folglich Wissen über den QID „`Geschlecht`“ und das SA „L-WBC“.

In Abschnitt 5.1.2.1 wurde festgelegt, dass die beteiligten Parteien eine kommutative Verschlüsselung auf Basis der XOR-Operation durchführen. Zu diesem Zweck benötigt jede Partei einen Schlüssel, der nur ihr bekannt ist und in jeder Iteration des Algorithmus wechselt [JC05, S. 173]. Wir wählen initial den Schlüssel $K_1 = 1$ für Partei 1 und den Schlüssel $K_2 = 2$ für Partei 2.

ID	PLZ	ALTER	KRANKHEIT
1	47677	29	Magengeschwür
2	47602	22	Gastritis
3	47678	27	Magenkrebs
4	47905	43	Gastritis
5	47909	49	Grippe
6	47906	47	Bronchitis
7	47605	30	Bronchitis
8	47673	36	Magenkrebs
9	47607	32	Lungenentzündung

(a) Attribute der Partei P_1

ID	GESCHLECHT	L-WBC
1	46,XY	11000
2	46,XY	10000
3	46,XY	9000
4	45,Xo	8000
5	46,XX	3000
6	46,XX	6000
7	47,XXY	7000
8	46,XY	5000
9	46,XY	4000

(b) Attribute der Partei P_2

ID	PLZ	ALTER	KRANKHEIT
1	476**	[20-29]	Magengeschwür
2	476**	[20-29]	Gastritis
3	476**	[20-29]	Magenkrebs
4	479**	[40-49]	Gastritis
5	479**	[40-49]	Grippe
6	479**	[40-49]	Bronchitis
7	476**	[30-39]	Bronchitis
8	476**	[30-39]	Magenkrebs
9	476**	[30-39]	Lungenentzündung

(c) 3-Anonymisierung von P_1

ID	GESCHLECHT	L-WBC
1	männlich	11000
2	männlich	10000
3	männlich	9000
7	männlich	7000
8	männlich	5000
9	männlich	4000
4	weiblich	8000
5	weiblich	3000
6	weiblich	6000

(d) 3-Anonymisierung von P_2

Tabelle 5.1: Vertikal partitionierte Daten eines Mehrparteien-Szenarios

Nehmen wir an die Parteien P_1 und P_2 einigten sich im Vorfelde der Anonymisierung auf ein k von 3 zum Schutz der Daten. Die VGHs seien wie in in Abbildung 3.1b, 3.1c und 3.1d gegeben.

Gemäß Zeile 1 des DPP$_2$GA-Protokolls berechnet jede Partei zunächst eine lokale k-Anonymisierung unter Zuhilfenahme des DF-Algorithmus[5]. Die jeweiligen Ergebnisse finden sich in Tabelle 5.1c und 5.1d.

Die Ergebnisse beider Parteien entsprechen bereits der Anonymisierung durch den DF-Algorithmus (vgl. Tabelle 4.1d). Wir werden nun beobachten wie beide Parteien den sicheren Vergleich der resultierenden Äquivalenzklassen vornehmen.

5 Die Wahl des DF-Algorithmus ist nicht vorgegebenen, es kommt jeder Algorithmus infrage mit dessen Hilfe eine valide k-Anonymisierung erzeugt werden kann.

Zunächst erstellt jede Partei ein Mengensystem γ_1^i aus den entstandenen Äquivalenzklassen. Es ergeben sich die folgenden Mengensysteme:

$$\gamma_1^1 = \{\{1,2,3\},\{4,5,6\},\{7,8,9\}\} \tag{5.6}$$
$$\gamma_1^2 = \{\{1,2,3,7,8,9\},\{4,5,6\}\} \tag{5.7}$$

Im nächsten Schritt gilt es die Mengensysteme zu vergleichen und gemäß Definition 32 auf Äquivalenz zu prüfen. Zu diesem Zweck ermitteln die Parteien jeweils das zweifach verschlüsselte Mengensystem $\Gamma_1^i, i \in \{1,2\}$, welches mittels kommutativer Verschlüsselung erzeugt wurde. Der erste Schritt der Verschlüsselung erfolgt elementweise über den Elementen der Teilmengen von $\gamma_1^i, i \in \{1,2\}$. Wir wollen diesen Schritt beispielhaft an der Verschlüsselung der Menge $\{1,2,3\}$ von Partei 1 nachvollziehen:

$$E_{K_1}(\{1,2,3\}) = \{1 \oplus 1, 2 \oplus 1, 3 \oplus 1\} = \{0,3,2\} \tag{5.8}$$

Insgesamt ergeben sich die folgenden verschlüsselten Mengensysteme:

$$E_{K_1}(\gamma_1^1) = \{\{0,3,2\},\{5,4,7\},\{6,9,8\}\} \tag{5.9}$$
$$E_{K_2}(\gamma_1^2) = \{\{3,0,1,5,10,11\},\{6,7,4\}\} \tag{5.10}$$

Diese können aufgrund ihrer Verschlüsselung über den unsicheren Kanal ausgetauscht werden. Beide Parteien verfügen nun über das unverschlüsselte Mengensystem ihrer eigenen Äquivalenzklassen sowie das verschlüsselte Mengensystem der Äquivalenzklassen der jeweilig anderen Partei. Ein direkter Vergleich der Mengensysteme ist hierdurch nicht möglich. Um dies zu erreichen, verwenden die Parteien Eigenschaft 1 der kommutativen Verschlüsselung. Sie erzeugen das zweifach verschlüsselte Mengensystem $\Gamma_1^i, i \in \{1,2\}$ wie folgt:

$$\begin{aligned}\Gamma_1^1 = E_{K_1}(E_{K_2}(\gamma_1^2)) &= E_{K_1}(\{\{3,0,1,5,10,11\},\{6,7,4\}\}) \\ &= \{\underbrace{\{2,1,0,4,11,10\}}_{\hat{=}\gamma_1^2[1]}, \underbrace{\{7,6,5\}}_{\hat{=}\gamma_1^2[2]}\}\end{aligned} \tag{5.11}$$

$$\begin{aligned}\Gamma_1^2 = E_{K_2}(E_{K_1}(\gamma_1^1)) &= E_{K_2}(\{\{0,3,2\},\{5,4,7\},\{6,9,8\}\}) \\ &= \{\underbrace{\{2,1,0\}}_{\hat{=}\gamma_1^1[1]}, \underbrace{\{7,6,5\}}_{\hat{=}\gamma_1^1[2]}, \underbrace{\{4,11,10\}}_{\hat{=}\gamma_1^1[3]}\} \quad .\end{aligned} \tag{5.12}$$

Das zweifach verschlüsselte Mengensystem $\Gamma_1^i, i \in \{1,2\}$ wird erneut über den unsicheren Kanal ausgetauscht. Beide Parteien verfügen nun über Γ_1^1 sowie Γ_1^2. Zeile 9 des Algorithmus 5.1 verlangt eine Prüfung der $\Gamma_1^i, i \in \{1,2\}$ auf Äquivalenz gemäß

Definition 32. Aufgrund der Kommutativitätseigenschaft ist es möglich die Mengensysteme zu vergleichen, ohne über deren konkrete Werte zu verfügen. Jede Partei stellt diesbezüglich ihre eigenen Beobachtungen an. Beide Parteien werden dieses Ergebnis infolge der HBC-Eigenschaft des Protokolls wahrheitsgemäß ermitteln. Unabhängig voneinander erlangen beide Parteien das folgende Ergebnis:

$$|\gamma_1^1[1] \cap \gamma_1^2[1]| = |\{2,1,0\}| \quad = 3 \geq k \tag{5.13}$$
$$|\gamma_1^1[2] \cap \gamma_1^2[2]| = |\{7,6,5\}| \quad = 3 \geq k \tag{5.14}$$
$$|\gamma_1^1[3] \cap \gamma_1^2[1]| = |\{4,11,10\}| = 3 \geq k \tag{5.15}$$

Es gilt die Äquivalenz der Mengensysteme nach Definition 32. Demgemäß ist es den Parteien nun gestattet die anonymisierten Teiltabellen T^1 und T^2 auszutauschen. Ein Join über dem Attribut `ID` liefert eine global k-anonyme Tabelle T_k. Das Resultat findet sich in schematischer Darstellung in Abbildung 5.3.

Algorithmus 5.1 : DPP$_2$GA

Eingabe : $P_i, T^i, QI_i, k, VGH_{A_m}, \forall A_m \in QI_i$ mit $i \in \{1,2\}$
Ausgabe : k-anonyme Tabelle $T_k = T_k^1 \bowtie_{\texttt{ID}} T_k^2$

1 P_i erzeugt lokale k-Anonymisierung durch Algorithmus 4.1
2 int cnt $\leftarrow 0$ `/* Anzahl der Iterationen */`
3 **repeat**
4 cnt $\leftarrow$ cnt $+ 1$
5 P_i erstellt γ_{cnt}^i `/* Erstellung Mengensystem */`
6 P_i erstellt $E_{K_i}(\gamma_{\text{cnt}}^i)$ und sendet es zu $P_j, j \neq i$
7 P_i empfängt $E_{K_j}(\gamma_{\text{cnt}}^j)$ und berechnet $\Gamma_{P_j} = E_{K_i}\left(E_{K_j}(\gamma_{\text{cnt}}^j)\right), j \neq i$
8 P_i empfängt $\Gamma_{P_i} = E_{K_j}\left(E_{K_i}(\gamma_{\text{cnt}}^i)\right), j \neq i$
9 **until** $\Gamma_1 \equiv \Gamma_2$
10 **return** $T_k \leftarrow T_k^1 \bowtie T_k^2$

5.2.2 *Analyse des DPP$_2$GA-Protokolls*

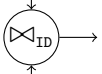
Beispiel 14 vermittelte uns ein Verständnis des Protokolls. Bislang unbeachtet blieb der Fall der Nicht-Äquivalenz der Mengensysteme. Jiang et al. sehen für diesen Fall keine explizite Behandlung im Protokoll vor. Vielmehr entspricht der Schritt der Erhöhung des Iterationszählers in Zeile 4 des Algorithmus einer erneuten k-Anonymisierung der bereits k-anonymen Daten. Durch diesen Schritt würde der Umfang

ID	PLZ	ALTER	KRANKHEIT
1	476**	[20-29]	Magengeschw.
2	476**	[20-29]	Gastritis
3	476**	[20-29]	Magenkrebs
4	479**	[40-49]	Gastritis
5	479**	[40-49]	Grippe
6	479**	[40-49]	Bronchitis
7	476**	[30-39]	Bronchitis
8	476**	[30-39]	Magenkrebs
9	476**	[30-39]	Lungenentz.

Partei 1 (P_1)

ID	GESCHL.	GEHALT
1	männlich	3000
2	männlich	4000
3	männlich	5000
7	männlich	7000
8	männlich	9000
9	männlich	10000
4	weiblich	6000
5	weiblich	11000
6	weiblich	8000

Partei 2 (P_2)



ID	PLZ	ALTER	GESCHL.	KRANKHEIT	GEHALT
1	476**	[20-29]	männlich	Magengeschw.	3000
2	476**	[20-29]	männlich	Gastritis	4000
3	476**	[20-29]	männlich	Magenkrebs	5000
4	479**	[40-49]	weiblich	Gastritis	6000
5	479**	[40-49]	weiblich	Grippe	11000
6	479**	[40-49]	weiblich	Bronchitis	8000
7	476**	[30-39]	männlich	Bronchitis	7000
8	476**	[30-39]	männlich	Magenkrebs	9000
9	476**	[30-39]	männlich	Lungenentz.	10000

Gemeinsame Veröffentlichung von P_1 und P_2

Abbildung 5.1: Resultat der verteilten k-Anonymisierung nach Jiang et al. [JC05]

der Äquivalenzklassen vergrößert. In der Folge wäre ein erfolgreicher Vergleich der Äquivalenzklassen nach Definition 32 wahrscheinlicher. Am Ende einer Generalisierungs-Hierarchie steht die Unterdrückung des Attributwerts. Hierdurch ist die Terminiertheit des Algorithmus gewährleistet, denn vollständig unterdrückte Attributwerte bilden eine Äquivalenzklasse vom Umfang der Einträge der Tabelle. Die IDs stimmen bei beiden Teiltabellen überein, woraus sich zwangsläufig die Äquivalenz ergibt.

Der Schritt der erneuten Anonymisierung birgt laut Jiang et al. Verbesserungspotential. Zur Erhöhung der Utility der Daten wäre es denkbar, die erneute k-Anonymisierung der Daten auf eine Partei zu beschränken [JC05, S. 176]. Hierdurch wird die Sicherheit des Verfahrens gefährdet. Der mehrfache Vergleich eines unveränderten Mengensystems könnte die Anonymisierung der Daten gefährden [JC05, S. 176]. Um diesem Umstand entgegen zu wirken, wäre ein Protokoll gemäß den Prinzipien der SMC notwendig. Dieses würde ausschließlich die Mächtigkeit der Schnittmenge ermitteln und ließe keine Rückschlüsse auf die an dem Schnitt beteiligten Elemente zu. Derartige SMC-Protokolle realisieren sogenannte *Privacy Preserving Set Operations* [KS05].

Ein weiteres Problem des DPP$_2$GA-Protokolls besteht in der abschließenden Vereinigung der beiden Teiltabellen durch den Join über dem Attribut ID. Dieser ermöglicht eine eindeutige Zuordnung von SAs zu IDs.

Zur Lösung der zuvor genannten Probleme entwickelten Jiang et al. ein erweitertes Verfahren Namens *DkA* [JC06]. Dieses bedient sich eines Protokolls zur sicheren Berechnung des Schnitts zweier Mengen (*Secure Set Intersection (SSI)*). Mit der Hilfe des SSI-Protokolls erfolgt der Vergleich der Äquivalenzklassen gemäß SMC-Vorgaben [JC06]. Weiterhin wird in diesem Verfahren ein sicherer Join mithilfe der kommutativen Verschlüsselung realisiert.

Protokolle der SMC erweisen sich oftmals als umfangreich [Ner+11]. Wir wollen daher an dieser Stelle lediglich vermerken, dass Protokolle zur Lösung der genannten Probleme existieren, ohne uns diesen im Detail zu widmen. Der Fortgang dieser Arbeit soll sich auf Konzepte und Probleme der Anonymisierung von verteilten Daten konzentrieren. Dabei steht die Nachvollziehbarkeit der Ergebnisse im Vordergrund. Die Optimalität der Ergebnisse sowie die Sicherheit gemäß den Vorgaben der SMC, sollen nachrangig behandelt werden.

Bevor wir uns dem Kern dieser Arbeit widmen, fassen wir unsere Erkenntnisse zusammen. Uns sind drei wesentliche Konzepte der Datenanonymisierung sowie deren Schwachstellen bekannt. Namentlich die k-Anonymisierung, die ℓ-Diversity sowie die t-Closeness. Wir haben ferner gesehen, wie eine k-Anonymisierung über verteilten Daten mithilfe des DPP$_2$GA-Protokolls realisiert werden kann. Dieses gebrauchte den uns bereits bekannten DF-Algorithmus.

Da eine k-Anonymisierung zum Erreichen einer t-Closeness wesentlich ist, wollen wir auf die Analyse von Protokollen zur verteilten Berechnung einer ℓ-Diversity verzichten. Wir werden in Kapitel 6 auf diese eingehen. Nachfolgend werden wir uns mit den Anforderungen beschäftigen, welche durch die Hinzunahme der Forderung der t-Closeness entstehen. Weiterhin soll uns beschäftigen, wie sich eine derartige Forderung in einem verteilten Szenario realisieren ließe.

5.3 t-CLOSENESS ÜBER VERTIKAL PARTITIONIERTEN DATEN

In dem vorhergehenden Abschnitt wurde gezeigt, wie sich das Konzept der k-Anonymisierung in einer verteilten Umgebung umsetzen lässt. Nach bestem Wissen des Autors existiert bislang kein Ansatz das Konzept der t-Closeness in einer verteilten Umgebung zu realisieren. Dies kann, neben dem schwindenden Interesse für die Konzepte des PPDP weitere Ursachen haben.

Während die Entwicklung und Verfeinerung von Konzepten zur Datenanonymisierung sowohl für den Bereich des PPDM als auch den Bereich des PPDP weitere

Fortschritte macht, erscheint im Zusammenhang mit der Anonymisierung von verteilten Daten ein bisher wenig beachtetes Problem: Der Schutz der Privatsphäre bei Existenz von mehr als einem SA, sogenannten Multiple-Sensitive-Attributes (MSA) [FAN11, S. 188]. Dieses Szenario beschränkt sich nicht nur auf das Konzept der t-Closeness, sondern findet sich bereits bei der ℓ-Diversity [GCG08]. Lokal operierende Konzepte, welche eine k-Anonymisierung herstellen, sind von solchen Einschränkungen nicht betroffen. Sie stellen keine Anforderungen an die sensiblen Attribute einer Veröffentlichung [GCG08; DB12].

Dies ändert sich bei einer verteilten k-Anonymisierung. Durch Erweiterung des Konzepts auf mindestens zwei Parteien, erhöht sich die Anzahl der SAs im selben Maße. Zusätzlich können Einschränkungen bezüglich des Vertrauens aller beteiligten Parteien untereinander vorherrschen. Es ergibt sich das Problem der *Background-Join-Attack* [FAN11]. Bei dieser Art von Angriff wird einem Angreifer ein gewisser Grad an Hintergrundwissen zugestanden. Im Gegensatz zu den in Kapitel 4 genannten Angriffen, bezieht sich dieses auf den konkreten Wert eines SA. Wendet der Beobachter dieses Wissen auf eine Veröffentlichung an, so kann er aus ihr weitere sensible Informationen extrahieren.

Sobald wir uns etwas näher mit den Anforderungen an ein verteiltes t-Closeness-Protokoll auseinander gesetzt haben, werden wir die Berechnung der t-Closeness in einem MSA-Szenario betrachten. Anschließend werden wir das Risiko einer Verletzung der Privatsphäre unter der Background-Join-Attack analysieren und den Grad der Gefahr aufzeigen. Den Kapitelabschluss bildet die Formulierung eines algorithmischen Konzepts, welches die t-Closeness von Daten in einer verteilten Umgebung sicherstellt.

5.3.1 *Anforderungen an ein verteiltes System*

Ein verteiltes System stellt eine Vielzahl von Anforderungen an die ausführende Umgebung. Aufgrund des dezentralen Charakters fällt besonderes Augenmerk auf die nicht-funktionalen Anforderungen.

Einleitend wurde erklärt, dass die vorliegende Diplomarbeit zum Ziel hat, ein Protokoll zur verteilten Veröffentlichung von Daten unter Einhaltung des Konzepts der t-Closeness zu entwickeln. Ein mögliches Anwendungsszenario wurde mit der Veröffentlichung von medizinischen Daten und deren Zusammenführung aus verschiedenen Studien angegeben (vgl. Kapitel 1). Wir wollen diese Rahmenbedingungen im Folgenden spezifizieren, um ein genaueres Bild dessen zu erhalten, was ein verteilter Datenanonymisierungsalgorithmus zu leisten imstande sein muss.

5.3.1.1 *Funktionale Anforderungen*

Das im Rahmen dieser Diplomarbeit entwickelte Protokoll, soll im Stande sein eine zwei-Parteien k-Anonymisierung nach dem Prinzip des DPP_2GA unter Einhaltung der t-Closeness zu ermöglichen. Grundlage ist eine FDG der QIDs nach dem DF-Algorithmus.

Bei den beteiligten Partien handelt es sich um medizinische Einrichtungen, welche Daten aus einer gemeinsamen Studie mithilfe eines dezentralen Systems zusammenführen. Dabei besitzen beide Parteien denselben Patientenstamm. Das bedeutet, dass jeder Patient der einen Datenbasis erscheint ebenfalls in der anderen Datenbasis. Weiterhin beschreibt jeder Eintrag in den Mikrodaten jeder Partei genau ein Individuum. Dies ist eine häufig getroffenen Annahme, da die Anonymisierungskonzepte k-Anonymisierung und ℓ-Diversity bei mehrfachem Auftreten eines Individuums in der Datenbasis nicht sicher sind [XTo6b]. Daraus folgt, dass die Parteien über gemeinsames Wissen bezüglich der Patienten verfügen. Namentlich sind dies die IDs. Diese können vielerlei Gestalt besitzen. Wir werden uns unter diesen eine fortlaufende Nummer vorstellen. Diese bezeichne pro Partei jeweils das gleiche Individuum. Weiterhin besitzt jede Partei Wissen über genau ein sensibles Attribut sowie eine Menge an Attributen, welche als QIDs dienen. In der Folge bildet die Verknüpfung der QIDs beider Parteien einen gemeinsamen QID. Einleitend wurde die Einschränkung des zu entwickelnden Konzepts auf ein SA formuliert. Das SA kann sowohl quantitativer als auch kategorialer Natur sein.

Aufgrund der Background-Join-Attack wird die Veröffentlichung der Daten in drei Tabellen partitioniert. Eine Tabelle enthält die QIDs sowie ein assoziierendes Attribut: Den Bucket-Identifikator (BID). Die beiden anderen Tabellen enthalten ebenfalls den BID sowie das jeweilige SA. Das assoziierende Attribut BID ordnet die sensiblen Attribute den entsprechenden QIDs zu.

Die partitionierte Veröffentlichung soll demselben Angreifermodell standhalten, welches von Jiang et al. sowie Li et al. benutzt wird. Da in diesem auch eine der beteiligten Parteien als Angreifer in Frage kommt, kann ein Missbrauch der sensiblen Daten durch die jeweiligen Institutionen nicht ausgeschlossen werden.

Die Kommunikation der Parteien muss daher über verschlüsselte Kanäle und mithilfe von SMC-Protokollen erfolgen. Wir attestieren den beteiligten Parteien ein Verhalten nach dem Grundsatz HBC.

Das Ziel der t-Closeness ist es, den Wissenszuwachs eines Angreifers zwischen einer maximal anonymisierten Veröffentlichung und der geplanten Veröffentlichung zu minimieren (vgl. Abschnitt 4.3.1). Aus diesem Grund besitzt ein Angreifer Hintergrundwissen über die Verteilung der SAs. Das zu konstruierende System muss daher den in Kapitel 4 vorgestellten Angriffen standhalten. Namentlich sind dies das

Datenverknüpfungsproblem, die Background Knowledge Attack, die Homogeneity Attack sowie die Skewness Attack und Similarity Attack.

Eine Randbemerkung zur Utility der anonymisierten Daten wird im weiteren Verlauf dieser Arbeit erfolgen, jedoch beansprucht der erarbeitete Algorithmus nicht ein optimales Resultat bezüglich Datenschutz und Utility zu erbringen. Vielmehr wird das Prinzip der Verständlichkeit und Nachvollziehbarkeit der einzelnen Schritte dem der Optimalität vorgezogen. Es ist jedoch unabdingbar, dass das Ergebnis des Algorithmus eine valide k-Anonymisierung unter Einbehaltung der t-Closeness-Eigenschaft liefert.

Die Schritte des Algorithmus sind so gewählt, dass sie mit minimalem Aufwand nachvollzogen werden können. Sie bauen auf dem Lesenden bekannten Algorithmen der vorherigen Kapitel auf.

5.3.1.2 *Nicht-funktionale Anforderungen*

Das Hauptaugenmerk des verteilten Protokolls liegt auf der Korrektheit des Resultats nach den in Abschnitt 5.3.1.1 gestellten Anforderungen.

Die verwendeten Algorithmen weisen bekannte Schwachstellen auf, welche sich in der Folge auf den zu erarbeitenden Algorithmus übertragen. Eine Beseitigung dieser Schwachstellen steht jedoch nicht im Fokus dieser Arbeit. Dies sei fortführenden Arbeiten durch die Adaption erweiterter Datenschutzkonzepte überlassen.

Diese Diplomarbeit widmet sich ausschließlich dem Gebiet des PPDP. Der Schwerpunkt liegt auf dem Schutz sensibler Daten durch Methoden der Anonymisierung mittels Generalisierung und Unterdrückung. Eine Abhandlung über Zuverlässigkeit und Ausfallsicherheit sowie die Sicherheit der Daten durch Zugriffskontroll-Mechanismen in einem DBMS ist nicht Teil dieser Arbeit. Wir wollen davon ausgehen, dass dieser Schutz besteht und die Mikrodaten durch Standard-Verschlüsselungsalgorithmen geschützt sind. Dies gelte weiterhin für den Austausch der Daten im Zuge des zu entwickelnden Protokolls. Sie seien insbesondere gegen Veränderung durch die Übertragung zwischen den Parteien geschützt. Diese Veränderungen können böswillig durch einen Angreifer oder unwissentlich durch physikalische Übertragungsfehler hervorgerufen werden. Das Erkennen von Übertragungsfehlern oder der Manipulation der übertragenen Daten ist im Entwurf des Protokolls nicht vorgesehen. Es sei an dieser Stelle erneut darauf hingewiesen, dass die beteiligten Parteien nach dem Prinzip HBC fungieren. Es ist durch die getroffenen Annahmen daher nicht mit einer mutwilligen oder technisch bedingten Veränderung der Daten zu rechnen.

Eine Frage, welche sich im Zuge der Diskussion um SMC-Protokolle stellt, ist die der Antwortgeschwindigkeit eines solchen Systems [Orl11; Ner+11]. Im Zuge dieser Arbeit wird diese Betrachtung vernachlässigt.

Es sei an dieser Stelle festgehalten, dass es sich bei dem Ergebnis des Algorithmus um die Anonymisierung einer einmaligen Veröffentlichung handelt. Das wiederkehrende Auftreten der Tupel bei erneuter Veröffentlichung liegt außerhalb des Fokus dieser Arbeit.

5.3.2 *Berechnung der t-Closeness für mehrere sensible Attribute*

Einleitend wurde auf das Problem sogenannter MSA hingewiesen. Bislang hat dieser Fall in Bezug auf die *t*-Closeness wenig Beachtung gefunden [FAN11, S. 189]. Die Autoren der ursprünglichen Veröffentlichung zur *t*-Closeness sind sich des Problems bewusst. Sie verweisen auf weitere Forschung hinsichtlich einer Kostenfunktion, welche mehrere Attribute berücksichtigt [LLV07, S. 9f]. Cao et al. weisen in ihrer Arbeit zum *t*-Closeness-Framework SABRE auf die Möglichkeit der Erweiterung auf den MSA-Fall hin [Cao+11, S. 80].

Im Folgenden wollen wir uns den Problemen stellen, welche eine Erweiterung bekannter Anonymisierungskonzepte um das Prinzip der *t*-Closeness mit sich bringt.

5.3.2.1 *Statistische Abhängigkeit sensibler Attribute*

In Kapitel 2 haben wir uns bereits mit dem Begriff der Abhängigkeit bzw. Unabhängigkeit von Merkmalen beschäftigt. Wir wollen das Problem an dieser Stelle genauer beleuchten. Es ergeben sich drei Möglichkeiten der statistischen Abhängigkeit:

1. Die betrachteten Attribute sind gemäß Definition 9 statistisch unabhängig
2. Die betrachteten Attribute sind gemäß Definition 17 perfekt statistisch abhängig
3. Es besteht eine statistische Abhängigkeit der Attribute, welche einen Wert im Intervall $(0, 1)$ annimmt

t-CLOSENESS FÜR STATISTISCH UNABHÄNGIGE ATTRIBUTE Beschäftigen wir uns mit dem ersten Fall: Der statistischen Unabhängigkeit. Betrachten wir hierzu Tabelle 5.2. Im Vergleich zu Tabelle 3.2 wurden die quasi-identifizierenden Attribute ausgeblendet. Der Fokus soll allein auf den Abhängigkeiten der SAs untereinander liegen. Wir wollen uns an dieser Stelle der Analyse zweier quantitativer sensibler Attribute zuwenden. Betrachten wir daher die SAs `Gehalt` und `L-WBC` hinsichtlich ihrer Abhängigkeit.

ID	GEHALT (*in €*)	L-WBC (Abs. Zahl/ml)	KRANKHEIT	CRP (*Referenzbereich*)
1	3000	11000	Magengeschwür	erhöht
2	4000	10000	Gastritis	stark erhöht
3	5000	9000	Magenkrebs	erhöht
4	6000	8000	Gastritis	erhöht
5	11000	3000	Grippe	normal
6	8000	6000	Bronchitis	leicht erhöht
7	7000	7000	Bronchitis	leicht erhöht
8	9000	5000	Magenkrebs	erhöht
9	10000	4000	Lungenentzündung	normal

Tabelle 5.2: Darstellung der sensiblen Attribute

In Kapitel 2 wurde auf die Möglichkeit des Erkennens eines statistischen Zusammenhangs durch das Betrachten der Kontingenztafel hingewiesen. Eine solche Darstellung findet sich in Tabelle 5.3.

	L-WBC									
GEHALT	3000	4000	5000	6000	7000	8000	9000	10000	11000	SUMME
3000	0	0	0	0	0	0	0	0	1/9	1/9
4000	0	0	0	0	0	0	0	1/9	0	1/9
5000	0	0	0	0	0	0	1/9	0	0	1/9
6000	0	0	0	0	0	1/9	0	0	0	1/9
7000	0	0	0	0	1/9	0	0	0	0	1/9
8000	0	0	0	1/9	0	0	0	0	0	1/9
9000	0	0	1/9	0	0	0	0	0	0	1/9
10000	0	1/9	0	0	0	0	0	0	0	1/9
11000	1/9	0	0	0	0	0	0	0	0	1/9
SUMME	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1

Tabelle 5.3: Kontingenztafel der Attribute `Gehalt` und L-WBC

Durch die abgebildete Häufigkeitsverteilung sind wir geneigt einen linearen Zusammenhang (vgl. Abschnitt 2.1.6) der sensiblen Attribute zu vermuten. Wir wollen dies mit den Methoden der Informationstheorie überprüfen. Die relativen Häufigkeiten, welche wir nach Abschnitt 2.2.1 als Wahrscheinlichkeiten interpretieren können,

entnehmen wir der Kontingenztafel. Diese weist $n = 9$ Einträge pro Attribut auf. Die Entropie in bit des Attributs `Gehalt` beträgt demnach:

$$\begin{aligned} H(\texttt{Gehalt}) &= -\sum_{i=1}^{9} P(x_i) \cdot \log_2\left(P(x_i)\right) = (-1) \cdot \frac{1}{9} \cdot \log_2\left(\frac{1}{9}\right) \cdot 9 \\ &= -\log_2\left(\frac{1}{9}\right) \quad . \end{aligned} \tag{5.16}$$

Nach den Rechenregeln für Logarithmen ergibt dies exakt

$$\frac{\ln 9}{\ln 2} \quad . \tag{5.17}$$

Das Ergebnis ist somit identisch zur maximalen Entropie

$$\log_2(n) = \log_2(9) = \frac{\ln 9}{\ln 2} \quad . \tag{5.18}$$

Selbiges gilt für die Entropie der `L-WBC`: $H(\texttt{L-WBC}) = {}^{\ln 9}/_{\ln 2}$. In der Folge ergibt sich eine Entropie der gemeinsamen Verteilung von

$$\begin{aligned} H(\texttt{Gehalt},\texttt{L-WBC}) &= -\sum_{i=1}^{9}\sum_{j=1}^{9} P(\texttt{Gehalt}_i,\texttt{L-WBC}_j) \log_2\left(P(\texttt{Gehalt}_i,\texttt{L-WBC}_j)\right) \\ &= \frac{\ln 9}{\ln 2} \quad . \end{aligned} \tag{5.19}$$

Wir erhalten einen wechselseitigen Informationsgehalt von

$$\begin{aligned} I(\texttt{Gehalt};\texttt{L-WBC}) &= H(\texttt{Gehalt}) + H(\texttt{L-WBC}) - H(\texttt{Gehalt},\texttt{L-WBC}) \\ &= \frac{\ln 9}{\ln 2} + \frac{\ln 9}{\ln 2} - \frac{\ln 9}{\ln 2} = \frac{\ln 9}{\ln 2} \quad . \end{aligned} \tag{5.20}$$

Und folglich die normierte Transinformation von

$$NI(\texttt{Gehalt};\texttt{L-WBC}) = \frac{I(\texttt{Gehalt};\texttt{L-WBC})}{\min(H(\texttt{Gehalt}), H(\texttt{L-WBC}))} = \frac{{}^{\ln 9}/_{\ln 2}}{{}^{\ln 9}/_{\ln 2}} = 1 \quad . \tag{5.21}$$

Wir folgern, dass die Attribute `Gehalt` sowie `L-WBC` perfekt statistisch abhängig sind. Demnach können wir aus der Existenz des einen Attributwerts die Existenz des anderen ableiten.

In diesem Zusammenhang wird von einer *Scheinkorrrelation* der Attribute gesprochen [Fah+07, S. 149]. Die Korrelation der Attributwerte ist zufälliger Natur und ein Einfluss des Gehalts auf die Anzahl der Leukozyten kausal nicht erklärbar.

Durch das Fehlen eines kausalen Zusammenhangs der Attribute ist es in diesem Fall nicht sinnvoll, für eine Anonymisierungs-Metrik wie der t-Closeness, die Beachtung beider Attribute zu verlangen. In der Folge könnten die Daten übermäßig stark anonymisiert werden. Für Daten dieser Art ist es daher zweckdienlich, die t-Closeness getrennt zu berechnen [FAN11, S. 192].

Nachfolgend wollen diesen Umstand für vollständig statistisch abhängige Attribute überprüfen.

t-CLOSENESS FÜR VOLLSTÄNDIG STATISTISCH ABHÄNGIGE ATTRIBUTE Wir wollen an dieser Stelle den kausalen Zusammenhang der Attribute vernachlässigen. Durch eine normierte Transinformation von 1 erweisen sich die Attribute `Gehalt` und `L-WBC` als vollständig statistisch abhängig. Aus der Abhängigkeit folgt, dass die Attribute eine identische t-Closeness für identische Äquivalenzklassen aufweisen würden [FAN11, S. 190]. Dies lässt sich leicht zeigen:

Theorem 3 (t-Closeness perfekt statistisch abhängiger Attribute)**.** *Seien die zwei Attribute X und Y perfekt statistisch abhängig, dann gilt für jede Teilmenge $Z \subseteq X$:*

$$d_{EMD}(Z \| X) = d_{EMD}(Z \| Y) \tag{5.22}$$

Beweis. Es gilt $\forall x_i \in X$ existiert genau ein $y_j \in Y$. Daraus folgt, dass $f(x_i) = f(y_j)$ für alle $x_i \in X, y_j \in Y$. Somit gilt nach Fahrmeir et al.: $X_X = X_Y$, da durch die relativen Häufigkeiten die empirische Verteilung von Daten vollständig beschrieben wird [Fah+07, S. 228]. Somit gilt für alle $Z \subseteq X : Z \subseteq X \implies Z \subset Y$. Hieraus folgt die Behauptung. □

t-CLOSENESS FÜR STATISTISCH ABHÄNGIGE ATTRIBUTE Es bleibt die Betrachtung des dritten Falls. Die Daten sind weder unabhängig noch vollständig statistisch abhängig. Jedoch besteht zwischen ihnen ein messbarer Zusammenhang und sogar eine Korrelation. Es existieren bislang keine Berechnungsgrundlagen für die EMD, welche die Auftrittswahrscheinlichkeit eines Attributwerts in Abhängigkeit eines anderen betrachten.

Wir wollen zur Analyse dieses Falls die sensiblen Attribute `Krankheit` sowie `CRP` heranziehen. Die relativen Häufigkeiten entnehmen wir der Kontingenztafel in Tabelle 5.4

Im Folgenden seien die Ergebnisse der Berechnung der Entropie, der gemeinsamen Entropie sowie der normierten und nicht-normierten Transinformation gegeben, ohne auf die Details ihrer Berechnung einzugehen – diese verläuft analog zum vorherigen Beispiel.

	CRP				
KRANKHEIT	NORMAL	LEICHT ERHÖHT	ERHÖHT	STARK ERHÖHT	$\sum$
MAGENGESCHW.	0	0	1/9	0	1/9
GASTR.	0	0	1/9	1/9	2/9
MAGENKR.	0	0	2/9	0	2/9
GRIPPE	1/9	0	0	0	1/9
LUNGENENTZ.	1/9	0	0	0	1/9
BRONCH.	0	2/9	0	0	2/9
$\sum$	2/9	2/9	4/9	1/9	1

Tabelle 5.4: Kontingenztafel der relativen Häufigkeiten von `Krankheit` und CRP

$$H(\texttt{CRP}) \approx 1,837 \tag{5.23}$$
$$H(\texttt{Krankheit}) \approx 2,503 \tag{5.24}$$
$$H(\texttt{Krankheit}, \texttt{CRP}) \approx 2,725 \tag{5.25}$$
$$I(\texttt{Krankheit}; \texttt{CRP}) \approx 1,614 \tag{5.26}$$
$$NI(\texttt{Krankheit}; \texttt{CRP}) \approx 0,879 \tag{5.27}$$

Wir stellen einen Zusammenhang der Attribute fest. Nach Li et al. ist es zweckdienlich die gemeinsame Verteilung der Attributwerte unter Verwendung einer gemeinsamen Ground-Distance zu errechnen [LLV07, S. 8]. Doch wie hoch ist die Aussagekraft einer t-Closeness über der gemeinsamen Verteilung? Zunächst würde eine Berechnungsvorschrift benötigt, welche eine gemeinsame t-Closeness ermittelt. Diese müsste zwei Attribute mit möglicherweise unterschiedlicher Skalierung berücksichtigen. Betrachten wir die Kontingenztafel 5.4 genauer, so stellen wir fest, dass die Ausprägung „normal“ des Attributs CRP zweimal vorhanden ist. Würden wir eine t-Closeness über der gemeinsamen Verteilung der beiden Attribute `Krankheit` und CRP berechnen, so könnten wir zwei *verschiedene* Ausprägungen des Werts „normal“ beobachten. Namentlich die Kombinationen {„Grippe“, „normal“} und {„Lungenentzündung“, „normal“}. In der Folge wandelt sich die Wahrscheinlichkeit des Auftretens des Attributwerts „normal“ von 2/9 zu 1/9. Dieses Phänomen ist als *Curse of Dimensionality* bekannt [Bel03].

Der Terminus Curse of Dimensionality beschreibt ein Problem, welches bei der Erhöhung der Dimension eines mathematischen Raumes entsteht. Steigt die Dimension des Raumes, so sind mehr Daten notwendig um diesen zu beschreiben.

Diese Problembeschreibung entspringt dem Bereich der Kombinatorik und multivariaten Statistik, für welchen sie ursprünglich von Bellman [Bel03] erfasst wurde. Anschaulich formuliert ist ein wesentlich größerer Stichprobenumfang erforderlich, um einen Raum höherer Dimension statistisch zu erfassen, als dies bei einer Dimension der Fall wäre. Im Bereich des PPDP wird mit der Dimension eines Eintrages je nach Anwendung eine Teilmenge seiner Attribute bezeichnet [Moh+10, S. 18:4]. Aggarwal et al. formulierten eine häufig zitierte Aussage, nach welcher der Abstand der Extrema zweier Verteilungen mit steigender Dimension der Daten gegen 0 konvergiert [AHK01, S. 422]. Somit sinkt die Aussagekraft von Metriken über einem höher-dimensionalen Raum. Aggarwal übertrug das Problem auf den Bereich des PPDP. In seiner Veröffentlichung postuliert Aggarwal eine übermäßige Vielfalt der Einträge mit wachsender Anzahl der QIDs [Agg05]. In der Folge wird das Erreichen einer k-Anonymisierung nur durch starke Generalisierung der Attributwerte möglich [Che+09, S. 145]. Durch die benötigte Anzahl an Generalisierungs- und Anonymisierungs-Schritten leidet die Utility der Daten [Moh+10, S. 18:4].

Fang et al. übertrugen das Problem auf die SAs einer Veröffentlichung [FAN11, S. 190]. Da aktuelle Anonymisierungskonzepte in der Mehrheit auf den Fall eines einzigen sensiblen Attributs zugeschnitten sind, liefern sie im MSA-Fall nicht die erwartete Sicherheit [FAN11, S. 189]. Durch die Existenz der Curse of Dimensionality ist die individuelle Berechnung der t-Closeness für jedes Attribut daher die einzige Möglichkeit diese auf den MSA-Fall zu übertragen [FAN11, S. 189].

Wir fassen unsere Erkenntnisse der vergangenen Abschnitte zusammen: Eingangs wurde festgestellt, dass die t-Closeness für nicht abhängige Attribute getrennt berechnet werden kann. Nachfolgend erweiterten wir unsere Feststellung auf vollständig statistisch abhängige Attribute. Letztendlich wurde im vorhergehenden Abschnitt das Problem der Curse of Dimensionality in Verbindung mit den sensiblen Attributen einer Veröffentlichung betrachtet.

Wir folgern daher in Übereinstimmung mit Fang et al., dass die attributweise Berechnung der t-Closeness zweckdienlich ist [FAN11, S. 189].

Im Zuge dieser Arbeit entstand eine Vorschrift für die Berechnung der Ground-Distance mehrerer sensibler Attribute. Sie kann als Verallgemeinerung der Ground-Distance von Li et al. angesehen werden und vereinfacht die attributweise Berechnung der t-Closeness für den Fall der MSAs. Für den weiteren Verlauf dieser Arbeit wird sie jedoch keine Verwendung finden. Dies liegt in der Struktur des Systems begründet, auf dessen Grundlage die Entwicklung eines verteilten Anonymisierungs-Protokolls fußt. Wir werden dies im folgenden Abschnitt näher betrachten. In Anhang A findet sich die Herleitung der Vorschrift. Des Weiteren wird dort auf die Schwierigkeiten der Entwicklung einer verallgemeinerten Vorgehensweise für mehrere kategoriale Attribute eingegangen.

5.3.3 *Neue Gefahren in einem verteilten t-Closeness-Protokoll*

Wir haben den Rahmen für ein verteiltes Anonymisierungs-Protokoll unter Berücksichtigung der *t*-Closeness-Eigenschaft gesetzt. Nun wollen wir uns mit den Risiken beschäftigen, welche durch die Verwendung eine solchen Protokolls entstehen, und der Frage, wie ihnen begegnet werden kann.

5.3.3.1 *Background-Join-Attack durch Multiple-Sensitive-Attributes*

Eingangs wurde erwähnt, dass der Join zweier anonymer Tabellen eine Background-Join-Attack ermöglicht. Dieses Szenario ergibt sich zwangsläufig in verteilten Umgebungen, in denen jede beteiligte Partei sensible Werte publizieren möchte.

Durch die Kenntnis der Mikrodaten, verfügt jede Partei über die Originaldaten der sensiblen Attribute ihrer Tabelle. Nach Veröffentlichung einer gemeinsamen Tabelle kann jede Partei die Werte ihrer sensiblen Attribute identifizieren, da diese nicht anonymisiert wurden. Über den Join der Tabellen erhalten die Parteien zusätzlich Informationen über die sensiblen Attribute der jeweilig anderen Partei. Dies stellt eine Verletzung der Privatsphäre dar.

Theorem 4 (Background-Join-Attack durch vertikale Partitionierung)**.** *Zwei Tabellen T^1 und T^2 mit jeweils einem sensiblen Attribut, SA_1 und SA_2, seien nach dem Prinzip Generalize-Then-Mashup in einer verteilten Umgebung anonymisiert. Dann gilt für die gemeinsame Tabelle $T_k = T_k^1 \bowtie T_k^2$: Die Wahrscheinlichkeit der Re-Identifikation eines Eintrages aus der anonymisierten Datenbasis ist größer als $1/k$. Das heißt, die Veröffentlichung genügt nicht der k-Anonymisierung für ein gegebenes k.*

Wir geben nachfolgend eine Intuition des Beweises.

Beweisidee. Nach den formulierten Anforderungen aus Abschnitt 5.3.1.1 fungieren alle beteiligten Parteien nach dem HBC Prinzip.

Vergegenwärtigen wir uns erneut Definition 24. Nach ihr ist ein Quasi-Identifikator eine Menge von Attributen $\{A_i, \ldots, A_j\} \subseteq \{A_1, \ldots, A_n\}$, welche mit externen Informationen verknüpft werden können und somit mindestens einen Eintrag ω_i aus $T(A_1, \ldots, A_n)$ eindeutig identifizieren.

Wäre das zu Grunde liegende Prinzip das des böswilligen Angreifers (vgl. Abschnitt 5.1), wäre der Beweis trivial: Der Angreifer würde die Werte seines sensiblen Attributs derart verändern, dass jedes Individuum durch diesen eindeutig zu identifizieren ist. In Anlehnung an Definition 24 würde gelten:

$$\forall \omega_i \in \mathcal{X} : f_g\left(\Pi_{SA_{T_k}}(f_c(\omega_i))\right) = \omega_i \quad . \tag{5.28}$$

Die Projektion auf das SA der Tabelle ermöglicht ein Re-Identifikation jedes Individuums. Dies entspricht dem sicheren Ereignis, dass der Wert eines sensiblen Attributs einem Eintrag zugeordnet werden kann. Verglichen mit der ursprünglichen Wahrscheinlichkeit von $1/k$, die besteht, wenn wir versuchen einem Eintrag den Wert eines sensiblen Attributs zuzuordnen, ist dies eine deutliche Erhöhung. Denn nach Korollar 2 gilt: $1/k < 1$. Somit findet eine Verletzung der Privatsphäre mit Wahrscheinlichkeit 1 statt.

Doch auch im Falle einer HBC-Umgebung ist die Anonymität der veröffentlichten Daten gefährdet. In jeder Äquivalenzklasse befinden sich nach Abschluss der k-Anonymisierung k Einträge. Diese weisen durch die Verwendung der ℓ-Diversity oder t-Closeness eine gewisse Verschiedenheit in den Attributwerten auf. Wir konzentrieren uns auf die t-Closeness-Eigenschaft.

Betrachten wir die Extremwerte der t-Closeness, beginnend mit einer t-Closeness von 0.

Eine Äquivalenzklasse ist 0-close, sobald die Einträge der Veröffentlichung dieselbe Verteilung der SAs aufweisen, wie die Einträge aller Äquivalenzklassen. Dies ist in zwei Fällen möglich:

1. Alle Einträge besitzen denselben sensiblen Attributwert
2. Jede Äquivalenzklasse enthält dieselbe Anzahl von mindestens zwei verschiedenen Attributwerten des sensiblen Attributs

Im ersten Fall würde von einer Veröffentlichung der Daten abgesehen werden, da selbst die erweiterten Konzepte wie die t-Closeness keinen Schutz böten.

Die einzige Möglichkeit ist daher Fall 2: Es existieren mindestens zwei verschiedene Attributwerte des sensiblen Attributs. Deren Anzahl ist aufgrund des t von 0 in jeder Äquivalenzklasse identisch. Sei diese Anzahl mit n gegeben. Es gilt:

$$n \in \mathbb{N} : 2 \leq n \leq k \quad . \tag{*}$$

Dann beträgt die Wahrscheinlichkeit einem Individuum einen sensiblen Attributwert zuzuordnen gerade n/k.
Wegen Korollar 2 und (*) gilt: $1/k < n/k$. Folglich ist die Äquivalenzklasse nicht mehr k-anonym.

Betrachten wir nun das andere Extremum der t-Closeness: Eine t-Closeness von 1.

In Kapitel 4 wurde erläutert, dass die EMD zwar Werte im Intervall $[0, 1]$ annehmen kann, dies aufgrund der Eigenschaften der t-Closeness für eben diese nicht gilt (vgl. Theorem 2). Für das Erreichen einer t-Closeness nahe 1 sind zwei Fälle denkbar.

1. Eine Äquivalenzklasse enthält nur Ausprägungen eines Attributwerts. Alle anderen Äquivalenzklassen enthalten nur Ausprägungen eines anderen Attributwerts

2. Eine Äquivalenzklasse enthält mindestens zwei sehr ähnliche Ausprägungen. Alle anderen Äquivalenzklassen enthalten nur Ausprägungen eines anderen unähnlichen Attributwerts

Im ersten Fall wäre aufgrund der Identität der Attributwerte eine eindeutige Zuordnung der Attributwerte möglich und somit eine Verletzung der Privatsphäre eingetreten.

Für den zweiten Fall ergibt sich, analog zum Fall der t-Closeness von 0, eine Anzahl von mindestens $n > 2$ Attributwertsausprägungen.

Somit gilt für beide Extrema, dass jede Äquivalenzklasse mindestens 2 Attributwertsausprägungen enthält.

Dies muss auch für alle $0 < t < 1$ gelten. Demnach ist die Anzahl verschiedener Attributwerte stets $n \geq 2$. Folglich beträgt die Wahrscheinlichkeit einem Individuum einen sensiblen Attributwert zuzuordnen stets n/k. Durch Eigenschaft (*) und Korollar 2 gilt daher für jedes $t : 1/k < n/k$.

Somit erweisen sich sämtliche Äquivalenzklassen in einer MSA-Umgebung als nicht mehr k-anonym. □

Ein Maß für die durch MSA entstehende Gefahr einer Background-Join-Attack findet sich in Form der *Information Exposure Ratio* in Fang et al. [FAN11, S. 194f]. Diese gebraucht die Transinformation zur Berechnung von Abhängigkeiten der sensiblen Attribute, aus welcher sich die Wahrscheinlichkeit des gemeinsamen Auftretens ableiten lässt.

Im Kontext dieser Arbeit ist uns bereits ein Beispiel begegnet, welches für eine Background-Join-Attack zugänglich ist.

Beispiel 15. Betrachten wir hierzu Abbildung 5.1. Nehmen wir an Partei 2 möchte die Werte des SA von Partei 1 erfahren. Nehmen wir weiter an, die Tabellen wurden mithilfe eines SMC-Protokolls vereinigt. Die in Abschnitt 5.2.2 genannte Schwachstelle kommt daher nicht zum Tragen. Da sämtliche Attributwerte für das SA `Gehalt` verschieden sind, kann Partei 2 die Veröffentlichung mit den ihr vorliegenden Daten verknüpfen. Die Folge ist eine Re-Identifikation sämtlicher Einträge.

Das Problem der MSA wurde 2007 von Machanavajjhala et al. erkannt und als *Multi-Attribute ℓ-Diversity* beschrieben [Mac+07, S. 23]. Die Lösung von Machanavajjhala et al. bestand darin, während der Berechnung der ℓ-Diversity eines jeden sensiblen Attributs die übrigen sensiblen Attribute als QIDs zu werten. Dass dieses

Vorgehen keinen ausreichenden Schutz bietet, wurde von Gal et al. [GCG08, S. 30] gezeigt.

Das verteilte Verfahren von Jiang et al. berücksichtigt das Problem der MSA nicht. Auch das Verfahren von Fung et al. widmet diesem Umstand keine Aufmerksamkeit, wenngleich es für die Verwendung von MSA konzipiert ist [Fun+11, S. 4].

Aktuelle Arbeiten von Das et al. und Fang et al. schlagen zur Verhinderung der Background-Join-Attack eine Trennung der QIDs und der SAs vor [DB12; FAN11]. Dieser Vorgang wird *Fragmentierung* genannt und zuweilen auch als *Partitionierung, Dekomposition, Bucketization* oder *Safe Grouping* bezeichnet. Prinzipiell spiegelt es eine nicht-verlustlose Zerlegung der Mikrodaten wieder (vgl. Kemper et al. [KE11, S. 179ff]). Es wird daher zuweilen auch als solche gekennzeichnet [Che+09, S. 146]. Die Zerlegung erfolgt in disjunkte Teiltabellen. Um einem Eintrag eine Menge von SAs zuordnen zu können, wird eine Verbindung über ein assoziatives Attribut hergestellt. Dieses repräsentiert die Äquivalenzklasse, in der sich der Eintrag befindet.

Wir halten diesen wichtigen Begriff fest und definieren nach Xiao et al. [XT06a, S. 141f]:

Definition 33 (Fragmentierung). *Sei T eine Tabelle mit Attributen $\{A_1, \ldots, A_m\}$. Die Menge der IDs sei mit $ID_T = \{A_a, \ldots, A_b\} \subset \{A_1, \ldots, A_m\}$ gegeben. T enthalte ferner eine Menge an QIDs $QI_T = \{A_c, \ldots, A_d\} \subset \{A_1, \ldots, A_m\}$. Weiterhin seien die SAs mit $SA_T = \{A_e, \ldots, A_f\} \subset \{A_1, \ldots, A_m\}$ bezeichnet. Die Mengen ID_T, QI_T sowie SA_T seien disjunkt. Insbesondere gelte $QI_T \cap SA_T = \emptyset$.*

Sei T_k eine k-Anonymisierung der Tabelle T. Das Schema von T_k wird um eine BID BID_T erweitert, welche die Äquivalenzklassen durch einen eindeutigen Identifikator kennzeichnet. Demnach gilt: $T_k = ID_T \cup QI_T \cup SA_T \cup BID_T$. Dann bezeichne eine Fragmentierung *von T die Zerlegung von T in die Teiltabellen $T_{QI} := \Pi_{QI_T, BID_T}(T)$ sowie $T_{SA} := \Pi_{BID_T, SA_T}(T)$.*

Durch diese Maßnahme ist eine direkte Assoziation der sensiblen Attributwerte mit den zugehörigen Individuen ausgeschlossen [XT06a, S. 142]. Durch die eindeutige Identifizierung des QID durch seine Originalwerte, bedarf es jedoch eines erweiterten Konzepts wie ℓ-Diversity oder t-Closeness, um die SA zu schützen [Che+09, S. 77].

Ursprünglich wurde dieses Verfahren für das Konzept der k-Anonymisierung unter Berücksichtigung der ℓ-Diversity erdacht [XT06a]. De Capitani di Vimercati et al. erweiterten das Konzept auf mehrere sensible Attribute [De +10]. Das et al. verfolgten das Ziel einer Übertragung auf das Konzept der (n, t)-Closeness [DB12, S. 10].

Freilich bleibt festzuhalten, dass keiner der genannten Autoren eine verteilte Umgebung betrachtet und die Verwendung eines fragmentierenden Algorithmus bislang noch aussteht.

Im Folgenden werden wir daher den Fall der Berechnung einer verteilten *t*-Closeness betrachten und Rahmenbedingungen aufzeigen, in denen eine solche gelingen kann. Abschließend soll eine mögliche Realisierung mittels eines bekannten Algorithmus skizziert werden, ohne die Methodik auf diesen Algorithmus einzuschränken.

5.3.4 *Entwurf eines verteilten DPP_2GA-Protokolls*

Wir wollen nun anhand uns bekannter Protokolle das Konzept der *t*-Closeness in einer verteilten Umgebung realisieren. Dazu betrachten wir zunächst die Schritte, welche ein verteiltes Protokoll durchlaufen muss, um zu einer validen Anonymisierung gemäß unserer Voraussetzung zu gelangen.

Wir fassen unsere Erkenntnisse zusammen: Uns ist nunmehr ein Protokoll bekannt, welches eine verteilte *k*-Anonymisierung auf Basis des DF-Algorithmus realisiert. Es wurde gezeigt, dass eine attributweise Berechnung der *t*-Closeness sinnvoll erscheint. Durch das Problem der Background-Join-Attack wurde uns die Tatsache bekannt, dass die SAs von den QIDs getrennt veröffentlicht werden müssen. Im Folgenden werden wir aus diesen Erkenntnissen ein verteiltes Protokoll zur *k*-Anonymisierung entwerfen, welches zusätzlich die Beachtung der *t*-Closeness-Eigenschaft gewährleistet.

Wir erinnern an dieser Stelle erneut an das Ziel dieser Arbeit. Gegenstand der Anonymisierung sind im Rahmen einer medizinischen Forschung dezentral erhobene Daten. Die beteiligten Institute verfügen über denselben Patientenstamm. Eine gemeinsame Anonymisierung und Veröffentlichung durch eine der Parteien oder eine TTP scheidet aus datenschutzrechtlichen Gründen aus (vgl. Kapitel 1). Durch den Umstand der identischen Einträge der jeweiligen Datenbasen, ermitteln die Parteien dieselbe Menge an quasi-identifizierenden Attributen. Sie verfügen jedoch über eine individuelle Menge an sensiblen Attributen, deren Veröffentlichung Ziel des Anonymisierungs-Vorgangs ist. Wir werden uns zunächst mit dem einfachsten Fall beschäftigen: Die Parteien verfügen nur über genau ein sensibles Attribut. Eine Erweiterung auf mehrere sensible Attribute wird in Kapitel 6 betrachtet. Durch diesen Fall entstehen Probleme, welche gesondert Betrachtung finden sollen. Zunächst wollen wir uns dem Entwurf des verteilten Protokolls zuwenden.

Widmen wir uns dem Aktivitätsdiagramm in Abbildung 5.2. Dieses stellt den Ablauf der Anonymisierung auf Seiten jeder beteiligten Partei dar.

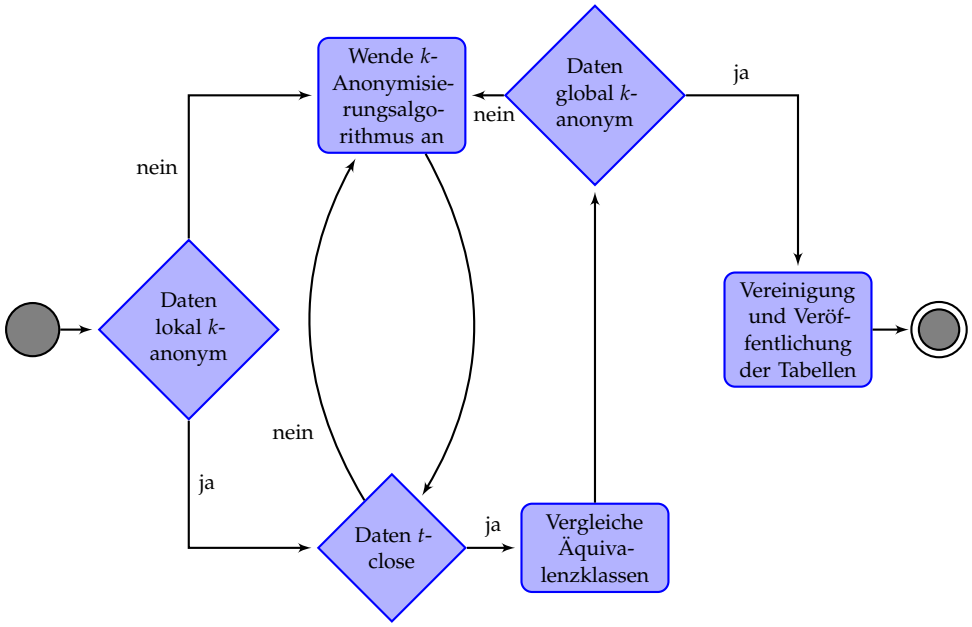


Abbildung 5.2: Aktivitätsdiagramm der verteilten k-Anonymisierung unter Beachtung der t-Closeness-Eigenschaft (vereinfachte Darstellung)

In Kapitel 4 wurde als Voraussetzung einer t-Closeness-Berechnung die Existenz von Äquivalenzklassen genannt. Anschließend wurde der Bezug dieser zur Gesamtveröffentlichung untersucht. Ein erweitertes verteiltes k-Anonymisierungs-Protokoll sollte eine ähnliche Vorgehensweise aufweisen und die Einhaltung der t-Closeness nach erfolgter k-Anonymisierung überprüfen. In der Folge sollten die entstandenen Äquivalenzklassen auf Konsistenz zwischen den Parteien geprüft werden. Im Falle einer Übereinstimmung können die QIDs und die SAs separat veröffentlicht werden. Im Falle eine Unterschieds müssen die Daten erneut generalisiert und das Prozedere wiederholt werden. Es kann an dieser Stelle darauf hingewiesen werden, dass eine erneute Prüfung auf t-Closeness nicht mehr notwendig ist. Dies ergibt sich aus der Tatsache, dass eine weitere Generalisierung einer Äquivalenzklasse, welche der t-Closeness genügt, höchstens t-close ist [Fakt 2 LLV07, S. 6]. Diese Eigenschaft wird zuweilen als *Monotonie* bezeichnet [BS08, S.].

Der Vorgang orientiert sich stark an dem Protokoll von Jiang et al. [JC05]. Dieses liefert bereits eine valide k-Anonymisierung von vertikal partitionierten Daten. Rufen wir uns Algorithmus 5.1 in Erinnerung. Zunächst wurde eine lokale k-Anonymisierung der Daten berechnet. DPP$_2$GA machte keine Vorgaben bezüglich des Anonymisierungs-Algorithmus und verwies der Einfachheit halber auf den DF-Algorithmus.

Anschließend wurden die entstandenen Äquivalenzklassen mithilfe eine SMC-Protokolls auf Äquivalenz geprüft. Im positiven Fall wurde von jeder beteiligten Partei ein Join der Daten gebildet und diese veröffentlicht. Im negativen Fall wurden die Daten lokal weiter generalisiert und die Prüfung wiederholt.

Dieses Prozedere erweitern wir um eine Prüfung der Äquivalenzklassen auf *t*-Closeness. Abschließend begegnen wir mit dieser Erweiterung von DPP_2GA der Background-Join-Attack durch *Fragmentierung*. Durch die Trennung der QIDs und SAs ist es zudem möglich, die quasi-identifizierenden Attribute in ihren original-Werten zu erhalten, da die *k*-Anonymisierung durch die Zerlegung sicher gestellt wird [XTo6a, S. 142]. Als finalen Schritt tauschen die beteiligten Parteien die Tabellen T^i_{QI} aus und bilden mit diesen eine Vereinigung über der ID. Anschließend werden die Tabellen $T^i_{QI} \bowtie T^j_{QI}$ und $T^i_{SA}, T^j_{SA}, i \neq j$ separat veröffentlicht.

Die Sicherheit der Anonymisierung werden wir im Anschluss an die Analyse des Protokolls zeigen. Diese bislang einzigartige Kombination dreier Konzepte ermöglicht, neben der Veröffentlichung verteilter Daten mithilfe aktuellster Anonymisierungskonzepte, eine Erhöhung der Utility der veröffentlichten Daten.

5.3.4.1 *Fragmenting t-Closeness-Enhanced DPP_2GA*

Wir wollen nachfolgend das Protokoll „Fragmenting *t*-Closeness-Enhanced DPP_2GA ($Ft EDPP_2GA$)“ diskutieren. Anschließend werden wir uns mithilfe einer Beispielrechnung seine Funktionsweise vergegenwärtigen.

Der Pseudocode ist Algorithmus 5.2 zu entnehmen. Im Unterschied zum 5.1-Protokoll, akzeptiert $Ft EDPP_2GA$ den Parameter *t* als Eingabe. Dieser gibt die für alle Äquivalenzklassen zu erreichende *t*-Closeness an.

Zunächst wird in Analogie zu DPP_2GA eine lokale *k*-Anonymisierung erzeugt. Wir wählen den DF-Algorithmus als Anonymisierungsalgorithmus aus denselben Gründen wie Jiang et al.: Der Algorithmus ist effektiv und leicht nachzuvollziehen [JCo5, S. 169]. Da es sich bei dem $Ft EDPP_2GA$-Protokoll um eine Erweiterung des DPP_2GA-Protokolls handelt, kommt jedoch jeder beliebige Algorithmus in Frage. In Zeile 2 wird das Resultat auf *t*-Closeness überprüft. Ist eine *k*-Anonymisierung gefunden, welche der *t*-Closeness genügt, so wird analog zu Algorithmus 5.1 verfahren. Fällt die *t*-Closeness-Berechnung negativ aus, so wird `Datafly` erneut angewandt.

Der Austausch der Äquivalenzklassen verläuft, unter Verwendung der kommutativen Verschlüsselung, analog zum DPP_2GA-Protokoll. Wir wollen uns daher nicht erneut in Details vertiefen und verweisen auf Abschnitt 5.2.1. Eine wichtige Änderung vollzieht sich jedoch in Zeile 13. Die von Jiang et al. definierte Äquivalenz-Relation kann bei der Forderung nach *t*-Closeness der Daten nicht verwendet wer-

Algorithmus 5.2 : FtEDPP$_2$GA

Eingabe : Tabellen T^i, Parteien $P_i, QI_{T^i}, k, t, VGH_{A_m} \forall A_m \in QI_{T^i}$ mit $i \in \{1,2\}$
Ausgabe : Tabelle $\Pi_{QI_T,BID}(T^i)$ der QIDs und Tabelle $\Pi_{BID,SA}(T^i)$ der SAs

```
1  float cost ← 1.0
2  while cost ≥ t do                                   /* Test auf t-Closeness  */
3      P_i erzeugt lokale k-Anonymisierung durch Algorithmus 4.1
4      foreach Äquivalenzklasse e ∈ E do
5          if SA quantitativ and d_EMD(e‖T_i) < cost then
6              cost ← Ergebnis von Algorithmus 4.2
7          else if SA kategorisch and d_EMD(e‖T_i) < cost then
8              cost ← Ergebnis von Algorithmus 4.3
9          end
10     end
11 end
12 int cnt ← 0
13 repeat
14     cnt ← cnt + 1
15     P_i erstellt γ^i_cnt
16     P_i erstellt E_{K_{P_i}}(γ^i_cnt) und sendet es zu P_j, j ≠ i
17     P_i empfängt E_{K_{P_j}}(γ^j_cnt) und berechnet Γ_{P_j} = E_{K_{P_i}}(E_{K_{P_j}}(γ^j_cnt)), j ≠ i
18     P_i empfängt Γ_{P_i} = E_{K_{P_j}}(E_{K_{P_i}}(γ^i_cnt)), j ≠ i
19 until Γ_{P_1} = Γ_{P_2}
20 P_i erstellt BID aus Äquivalenzklassen
21 return Π_{QI_T,BID}(T^i), Π_{BID,SA}(T^i)
```

den. Dem Grund werden wir uns in Abschnitt 5.3.4.2 widmen. Wir fordern an dieser Stelle die Gleichheit der Mengensysteme:

Definition 34 (Gleichheit zweier Mengensysteme). *Seien $\gamma^i_\alpha, \gamma^j_\beta$ zwei Mengensysteme, d. h. zwei Mengen welche Mengen enthalten. Bezeichne ferner $\gamma^i_\alpha[p]$ die p-te Teilmenge von γ^i_α und $\gamma^j_\beta[q]$ bezeichne die q-te Teilmenge γ^j_β. Zwei Mengensysteme $\gamma^i_\alpha, \gamma^j_\beta$ sind gleich genau dann, wenn die in ihnen enthaltenen Mengen gleich sind, d. h. es gilt:*

$$\forall \gamma^i_\alpha[p] \exists \gamma^j_\beta[q] : \gamma^i_\alpha[p] = \gamma^j_\beta[q] \quad .$$

Wir schreiben $\gamma^i_\alpha = \gamma^j_\beta$.

Die Besonderheit des FtEDPP$_2$GA-Protokolls und ein entscheidender Unterschied zum DPP$_2$GA-Protokoll, liegen in der Veröffentlichung der Daten begründet. Es wurde erläutert, dass diese fragmentiert zu erfolgen hat. Zeile 20 liefert den Grundstein hierzu. Sind die Äquivalenzklassen der Parteien identisch, so erstellt jede Partei ein assoziatives Attribut – den sogenannten BID. Dieser enthält eine ID, welche die Äquivalenzklasse des Attributwerts kennzeichnet. Die beteiligten Parteien können nun die Tabellen, welche die QIDs und den BID enthalten, durch einen Equi-Join über dem Attribut ID vereinen. Dies ist möglich, da die QIDs nach Definition keine sensiblen Daten enthalten. Abschließend veröffentlicht jede Partei zwei Tabellen. Zum einen die gesammelten *Originaldaten* der QIDs, verbunden mit dem BID. Zum anderen ihr sensibles Attribut, welches ebenfalls um den BID ergänzt wurde.

Sind die Äquivalenzklassen nicht identisch, so erstellen beide Parteien eine weitere Generalisierung ihrer Daten. Dieser Schritt ist nicht explizit Bestandteil des Algorithmus 5.2. Wie schon in Algorithmus 5.1, ergibt er sich implizit aus der Erhöhung des Zählers cnt und der damit verbundenen Änderung des Mengensystems γ^i (vgl. Abschnitt 5.2.2). Der Vergleich der Mengensysteme wird wiederholt, bis sich die Identität der Äquivalenzklassen einstellt.

Nachdem die Funktionsweise des Protokolls verdeutlicht wurde, stellen wir fest, dass sich dieses in vier Phasen unterteilen lässt. Die Abgrenzung der einzelnen Schritte ist wesentlich für die Überlegungen zur Korrektheit des Protokolls, wie wir sie im nachfolgenden Abschnitt anstellen werden. Wir halten daher das Folgende fest:

Definition 35 (Die vier Phasen des FtEDPP$_2$GA-Protokolls)**.**

1. *k-Anonymisierung*
2. *Validierung der t-Closeness*
3. *Berechnung gemeinsamer Äquivalenzklassen*
4. *Fragmentierung und Veröffentlichung*

Bevor wir uns der Diskussion des Protokolls zuwenden, wollen wir uns die Funktionsweise des Protokolls anhand eines Beispiels verdeutlichen.

Beispiel 16. Betrachten wir erneut ein Zwei-Parteien-Szenario, wie es in Tabelle 5.1 beschrieben ist. Partei 1 verfüge demnach über die QIDs „`PLZ, Alter`“ sowie das SA `Krankheit`. Partei 2 obliege das Wissen über den QID „`Geschlecht`“ und das SA `L-WBC`. Analog zu Beispiel 14 verwenden die Parteien eine kommutative Verschlüsselung auf Basis der XOR-Operation zur Bestimmung der Gleichheit der Mengensysteme.

Wir wählen initial den Schlüssel $K_1 = 1$ für Partei 1 und den Schlüssel $K_2 = 2$ für Partei 2.

Nehmen wir an die Parteien P_1 und P_2 einigten sich auf die folgenden Parameter: Es seien ein k von 3 sowie ein t von $0,3$ zum Schutze der Daten vereinbart. Die VGHs seien erneut wie in in Abbildung 3.1b, 3.1c und 3.1d gegeben.

Gemäß Zeile 3 des Algorithmus berechnet jede Partei zunächst eine lokale k-Anonymisierung unter Zuhilfenahme des DF-Algorithmus. Das Resultat dieser Berechnung ist identisch mit dem aus Beispiel 14. Die lokalen Ergebnisse können den Tabellen 5.1c-5.1d entnommen werden. Entsprechend finden sich parteiübergreifend die Äquivalenzklassen $E_1 = \{1,2,3\}, E_2 = \{4,5,6\}$ sowie $E_3 = \{7,8,9\}$. Es gilt die Konformität zur t-Closeness zu überprüfen. Zu diesem Zweck seien nachfolgend die Ergebnisse der t-Closeness-Berechnung gegeben, ohne auf die Details ihrer Berechnung eingehen zu wollen. Diese können Kapitel 4 entnommen werden.

ID	PLZ	ALTER	KRANKHEIT
1	476**	[20-39]	Magengeschwür
2	476**	[20-39]	Gastritis
3	476**	[20-39]	Magenkrebs
7	476**	[20-39]	Bronchitis
8	476**	[20-39]	Magenkrebs
9	476**	[20-39]	Lungenentzündung
4	479**	[40-59]	Gastritis
5	479**	[40-59]	Grippe
6	479**	[40-59]	Bronchitis

Tabelle 5.5: Wiederholte Anwendung des DF-Algorithmus

Attribut `Krankheit` von Partei 1:

$$d_{EMD}(E_1 \| V) = {}^{12}/_{27} \approx 0,444 \tag{5.29}$$
$$d_{EMD}(E_2 \| V) = {}^{8}/_{27} \approx 0,296 \tag{5.30}$$
$$d_{EMD}(E_3 \| V) = {}^{6}/_{27} \approx 0,222 \tag{5.31}$$

Attribut L-WBC von Partei 2:

$$d_{EMD}(E_1 \| V) = {}^{6}/_{72} \approx 0,083 \tag{5.32}$$
$$d_{EMD}(E_2 \| V) = {}^{11}/_{72} \approx 0,153 \tag{5.33}$$

Das Attribut `Krankheit` erweist sich demnach als $0,444$-close. Das Attribut `L-WBC` als $0,153$-close. Gemäß Zeile 2 ergibt der Vergleich von cost und t auf Seiten der Partei

1: $0,444 > 0,3$. Die Anonymisierung genügt also nicht der geforderten t-Closeness. Partei 1 ist gezwungen den DF-Algorithmus erneut anzuwenden. Das Ergebnis findet sich in Tabelle 5.5. Die Vereinigung der Äquivalenzklassen E_1 und E_3 zur neuen Klasse E_{13} weist eine t-Closeness von

$$d_{EMD}(E_{13} \| V) = 4/27 \approx 0,148 \tag{5.34}$$

auf.

Alle Äquivalenzklassen genügen der Forderung einer t-Closeness von höchstens $0,3$. Gemäß des DPP_2GA-Protokolls aus Abschnitt 5.2.1 tauschen die Parteien P_1 und P_2 ein verschlüsseltes Mengensystem γ_1^i aus. Der Übersichtlichkeit halber verzichten wir auf die Darstellung des Austauschs der verschlüsselten Daten, diese kann Beispiel 14 entnommen werden. Betrachten wir die durch die Äquivalenzklassen erzeugten Mengensysteme:

$$\gamma_1^1 = \{\{1,2,3,7,8,9\},\{4,5,6\}\} \tag{5.35}$$
$$\gamma_1^2 = \{\{1,2,3,7,8,9\},\{4,5,6\}\} \tag{5.36}$$

Es gilt die Gleichheit der Mengensysteme nach Definition 34. Gemäß Zeile 20 aus Algorithmus 5.2 erstellen die beteiligten Parteien einen BID für die Äquivalenzklassen und fügen diesen in die zu veröffentlichenden Tabellen ein. Das Resultat findet sich in schematischer Darstellung in Abbildung 5.3.

5.3.4.2 *Diskussion des Protokolls*

Im vorhergehenden Abschnitt wurde eine algorithmische Lösung des Problems der verteilten k-Anonymisierung unter Beachtung der t-Closeness diskutiert. Das adaptierte Protokoll DPP_2GA von Jiang et al. liefert eine nachweislich korrekte k-Anonymisierung [JC05, S. 174]. Es bleibt zu zeigen, dass diese Eigenschaft in dem Protokoll $FtEDPP_2GA$ erhalten bleibt. Da die Schritte der k-Anonymisierung, Berechnung der t-Closeness sowie der Prüfung der Äquivalenzklassen und Veröffentlichung der Tabellen streng voneinander getrennt sind, genügt es die Sicherheit der einzelnen Komponenten zu prüfen.

Hierzu betrachten wir zunächst eine entscheidende Änderung. Im vorangegangenen Abschnitt wurde die Äquivalenz-Relation zweier Mengensysteme zu einer Gleichheits-Relation verschärft. Dieser Schritt ist notwendig, da ansonsten die t-Closeness der gemeinsamen Veröffentlichung nicht gewährleistet werden kann.

Folgende Überlegung verdeutlicht das Gesagte:

Theorem 5. *Definition 34 ist wesentlich zur Wahrung der t-Closeness im Kontext des $FtEDPP_2GA$-Protokolls.*

Beweisidee. Nehmen wir an eine Partei benötige zum Erreichen der t-Closeness eine Äquivalenzklassen der Größe $2k$. Diese setze sich aus zwei Äquivalenzklassen der anderen Partei zusammen. Der anderen Partei gelänge es die t-Closeness für das vereinbarte k zu erhalten. Dann besäße gemäß Definition 32 der Schnitt jeder Teilmenge von γ_1^1 mit einer Teilmenge von γ_1^2 die Mächtigkeit k. Der Algorithmus wäre erfolgreich, die Ergebnisse würden vereint. Das Resultat wäre k-anonym, nicht jedoch t-close.

Dies begründet sich durch die verschiedenen Attributwerte der QIDs. Die Äquivalenzklasse der Mächtigkeit $2k$ würde in zwei Klassen der Größe k aufgeteilt werden. Demnach würde die t-Closeness-Eigenschaft verletzt, denn wir stellten zuvor fest, dass diese für eine Äquivalenzklasse nur für eine Mächtigkeit von $2k$ zu erreichen ist. □

Widmen wir uns nun, der Korrektheit der adaptierten Äquivalenz-Relation.

Theorem 6 (Korrektheit von FtEDPP$_2$GA)**.** *Definition 34 ist ein Spezialfall von Definition 32. Hierdurch bleibt die Korrektheit des Protokolls unberührt. Das heißt, das Ergebnis ist eine korrekte k-Anonymisierung unter Berücksichtigung der t-Closeness.*

Beweis. Wir zeigen dies, indem wir nachweisen, dass es sich bei der Gleichheits-Relation um einen Spezialfall der Äquivalenz-Relation handelt, für welche die Sicherheit schon bewiesen wurde [JC05, S. 174].

Zeigen wir zunächst die Einhaltung der k-Anonymisierung. Nach Definition 32 sind zwei Mengensysteme äquivalent, sobald der Schnitt jeder Teilmenge eines Mengensystems mit einer Teilmenge des anderen Mengensystems die Mächtigkeit größer oder gleich k hat. Definition 34 fordert die Identität der Mengensysteme. Da die Mengensysteme lokal k-anonym sind, sind sie nach Definition 34 auch global k-anonym. Folglich ist jeder Schnitt der Teilmengen größer als k. Somit gilt die Behauptung.

Es bleibt zu zeigen, dass die Eigenschaft der t-Closeness unberührt bleibt. In Zeile 3 des FtEDPP$_2$GA-Protokolls wird eine lokale k-Anonymisierung erreicht. Diese wird auf t-Closeness überprüft. Da es sich bei dem Anonymisierungsalgorithmus um den DF-Algorithmus handelt, wird in endlicher Anzahl an Schritten eine totale Unterdrückung sämtlicher Attributwerte der QIDs erlangt. Die einzig entstehende Äquivalenzklasse ist diejenige, welche sämtliche Tupel enthält. Folglich sind die Verteilungen der SAs der Äquivalenzklasse und der gesamten Veröffentlichung identisch. Die t-Closeness beträgt demnach 0. Dies ist der minimale Wert den das t annehmen kann. Da die Äquivalenzklassen beider Parteien identisch sein müssen, wird demnach die t-Closeness-Eigenschaft nicht verletzt. Ist keine Totalunterdrückung der Werte notwendig um eine t-Closeness zu erreichen, so bleibt die t-Closeness-

Eigenschaft durch das Monotonie-Verhalten der t-Closeness bei weiteren Generalisierungsschritten unberührt (vgl. Abschnitt 5.3.4). □

Da die Schritte der Berechnung gleicher Äquivalenzklassen in Zeile 13 bis 19 zu denen des DPP$_2$GA-Protokolls identisch sind, findet durch die Verwendung der kommutativen Verschlüsselung keine Verletzung der Privatsphäre statt. Dies wurde in Jiang et al. [JC05, S. 174ff] gezeigt. Xiao et al. zeigten zudem, dass durch die Fragmentierung der Daten keine Verletzung der Privatsphäre stattfindet [XT06a, S. 142].

Wir können daher festhalten, dass das FtEDPP$_2$GA-Protokoll eine korrekte k-Anonymisierung unter Wahrung der t-Closeness liefert. Gleichsam ist die Privatsphäre der Individuen, die durch die Daten repräsentiert werden, durch das kryptografische Protokoll unter denselben Bedingungen geschützt, die auch für das DPP$_2$GA-Protokoll gelten (vgl. Abschnitt 5.2.2).

Die Fragmentierung erweitert das DPP$_2$GA-Protokoll zudem um ein entscheidendes Detail: Die Veröffentlichung der Originaldaten. Durch die Fragmentierung erlangt FtEDPP$_2$GA eine höhere Utility der Daten als dies durch Generalisierung der Fall wäre, da die Originalwerte der QIDs erhalten bleiben [DB12, S. 4]. Dieser Utility-Gewinn auf Seiten der QIDs birgt jedoch einen Nachteil: Durch die Fragmentierung wird nicht nur der Zusammenhang zu den QIDs gelöst, sondern ebenfalls der Zusammenhang der SAs untereinander [FAN11, S. 194]. Fang et al. entwickelten die *Association Loss Ratio*, um den Informationsverlust zu beziffern [FAN11, S. 194].

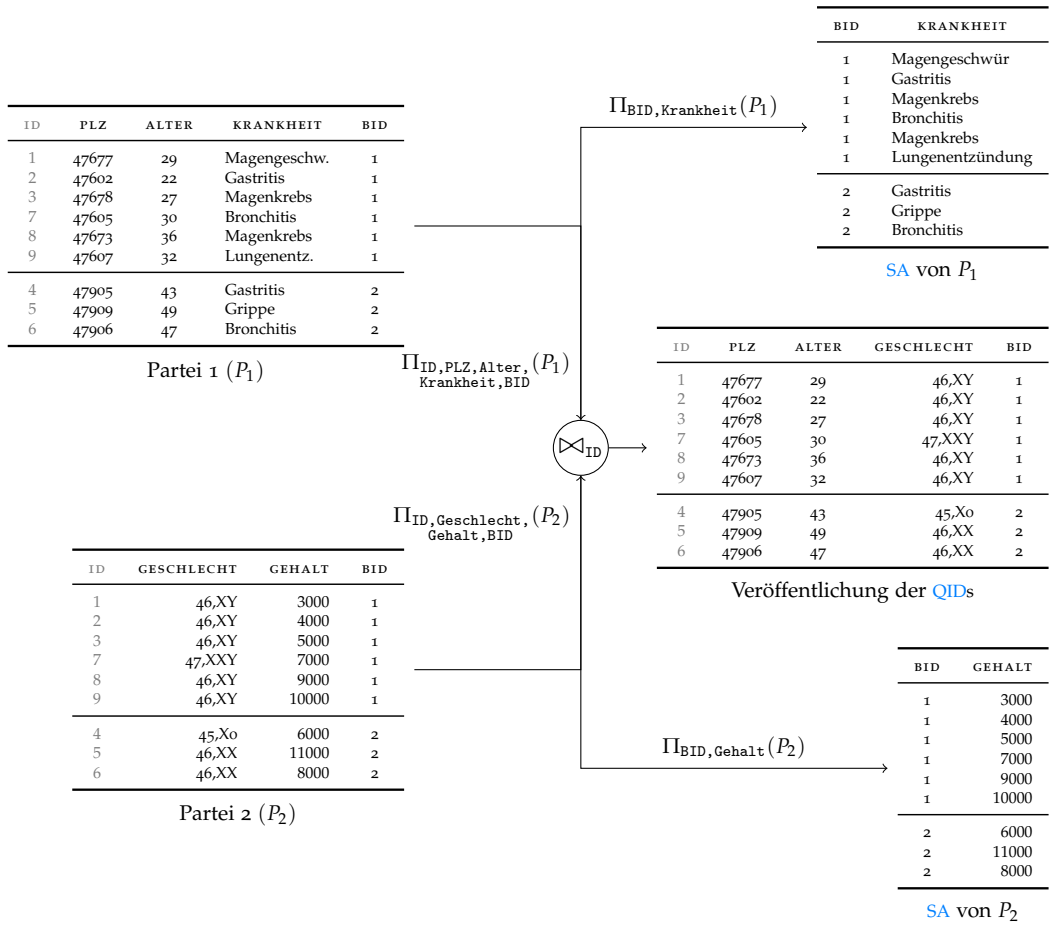


Abbildung 5.3: Resultat der verteilten k-Anonymisierung unter Beachtung der t-Closeness-Eigenschaft

6 ZUSAMMENFASSUNG UND AUSBLICK

Im Kontext der vorliegenden Diplomarbeit wurden Konzepte des Privacy-Preserving Data Publishing (PPDP) hinsichtlich ihrer Anwendbarkeit in einem verteilten Szenario untersucht. Besonderes Augenmerk lag auf der Betrachtung des Konzepts der t-Closeness, da für dieses bislang keine Erweiterung auf ein verteiltes Szenario bekannt ist. Für die Realisierung einer dezentralen Datenanonymisierung mithilfe der t-Closeness wurde untersucht, inwieweit sich vorhandene Methoden der Anonymisierung von vertikal partitionierten Daten um die Eigenschaft der t-Closeness erweitern lassen.

In Abschnitt 6.1 werden wir die Ergebnisse der Arbeit zusammenfassen. In diesem Zusammenhang werden wir die getroffenen Annahmen erläutern, welche für ein verteiltes Szenario wesentlich sind.

Einen Überblick zu anderen Arbeiten im Fachgebiet des PPDP wird Abschnitt 6.2 geben.

Abschnitt 6.3 wird diese Resultate zur aktuellen Forschung im Bereich der verteilten Anonymisierung von Daten in Relation setzen. Zudem wird in diesem Abschnitt analysiert werden, in welcher Hinsicht eine Optimierung aktueller Methoden erfolgen kann.

6.1 ERGEBNISSE

Im Zuge der Arbeit wurde ein Überblick über Methoden des PPDP erarbeitet. Zunächst wurde das grundlegende Konzept der k-Anonymisierung formal dargelegt. In diesem Zusammenhang haben wir uns den Begriff der Äquivalenzklassen erschlossen, auf welchem viele Konzepte der Datenanonymisierung basieren. Es wurde anhand der einfachen Heuristik des Datafly-Algorithmus (DF-Algorithmus) erläutert, wie derartige Äquivalenzklassen erstellt werden können. Nachfolgend wurde die Methodik der ℓ-Diversity mit ihren Ausprägungen der Distinct ℓ-Diversity sowie der Entropy ℓ-Diversity entwickelt. Dieser Schritt wurde durch betrachtete Schwächen des Konzepts der k-Anonymisierung motiviert. Das Konzept der ℓ-Diversity verlangt eine Mannigfaltigkeit der sensiblen Attributwerte innerhalb jeder Äquivalenzklasse. Eine Betrachtung der Semantik der Attributwerte findet jedoch nicht statt. Aus diesem Grund entstand die Notwendigkeit der Entwicklung weiterführender

Konzepte des Datenschutzes. In diesem Zusammenhang wurde das viel beachtete Konzept der t-Closeness im Detail besprochen.

Als erste Entwicklung im Bereich der Datenanonymisierung versucht sich das Konzept der t-Closeness an der Quantifizierung des Abstands zweier Attributwerte. Grundlage hierzu bildet das Maß der Earth-Movers-Distanz (EMD). Wir haben gezeigt, dass es sich bei der EMD um eine Ausprägung des Transportproblems handelt, welches mittels Methoden der linearen Optimierung gelöst werden kann. Das entscheidende Element der Distanzbestimmung bildet die Häufigkeitsverteilung der sensiblen Attributwerte einer Äquivalenzklasse. Durch die Betrachtung der Häufigkeitsverteilungen ergibt sich das Problem unterschiedlicher Skalenniveaus. Aus diesem Grund definiert die EMD verschiedene Metriken als Distanzmaß. Zum einen die Ordered Distance für quantitative Attributwerte, zum anderen die Hierarchical Distance für kategoriale Attributwerte. Die detaillierte Betrachtung der Metriken war für den nächsten Abschnitt der Arbeit von besonderer Wichtigkeit.

Die zentrale Forderung der Arbeit bestand darin, das Konzept der t-Closeness hinsichtlich seiner Eignung für ein verteiltes Szenario zu untersuchen. Dieser Schritt wurde durch die Existenz multizentrischer medizinischer Studien begründet. Wir konnten sehen, dass derartige Methoden für das Konzepte der k-Anonymisierung existieren.

Die k-Anonymisierung der Daten bildet die Grundlage für die Anwendung der t-Closeness. Daher bildet ein Protokoll zur k-Anonymisierung von vertikal partitionierten Daten, die Voraussetzung für das Überführen der t-Closeness in eine verteilte Umgebung. Ein derartiges Protokoll stellt das „Distributed Privacy-Preserving two-Party Generic Anonymizer“ Protokoll (DPP_2GA) dar. Dieses Protokoll realisiert die wohlbekannte Heuristik des DF-Algorithmus in einem verteilten Szenario.

In der Folge konnten Gründe identifiziert werden, weshalb das Konzept der t-Closeness bislang nicht in einem verteilten Szenario realisiert wurde. Es ergab sich das Problem der Multiple-Sensitive-Attributes (MSA). Dieses Problem fand in der Literatur bislang wenig Beachtung. Die t-Closeness errechnet sich auf Grundlage der Beziehungen der sensiblen Attributwerte zueinander. Durch die Existenz mehrerer sensibler Attribute entstand die Frage, in welcher Beziehung die Attributwerte unterschiedlicher Attribute stehen. Zu diesem Zweck wurden die Metriken Ordered Distance sowie Hierarchical Distance hinsichtlich ihrer Erweiterung auf mehrere sensible Attribute überprüft. Dieser Versuch erwies sich als nicht zielführend.

Für kategoriale Attribute wurde anhand eines Beispiel gezeigt, dass eine Verallgemeinerung der Hierarchical Distance auf mehrere sensible Attribute nicht möglich ist. Für quantitative Merkmale wurde eine Verallgemeinerung der Ordered Distance konzipiert: Die MSA-Ordered Distance. Diese Überlegungen finden sich in Anhang A. Die entstandene Metrik berücksichtigt nicht die statistische Abhängigkeit der sensi-

blen Attribute. Es wurden daher Überlegungen angestellt, ob die Berücksichtigung von statistischen Abhängigkeiten bei der Berechnung der t-Closeness relevant ist. Es konnte gezeigt werden, dass es aufgrund des Phänomens der Curse of Dimensionality nicht sinnvoll ist, die t-Closeness über der gemeinsamen Häufigkeit von Attributwerten zu berechnen.

Ausgehend von dieser wichtigen Feststellung, wurde ein kryptographisches Protokoll konzeptionell erarbeitet. Das entstandene Protokoll „Fragmenting t-Closeness-Enhanced DPP_2GA“ ($Ft EDPP_2GA$) ermöglicht die Berechnung einer k-Anonymisierung unter Beachtung der t-Closeness. Aufgrund der Eingangs definierten Forderung nach der Anschaulichkeit des Verfahrens, wurde die aus Kapitel 4 bekannte Erweiterung des DF-Algorithmus – DPP_2GA – als Grundlage des Verfahrens gewählt. Durch das definierte Szenario einer multizentrischen Studie und bekannter Schwächen des DPP_2GA-Protokolls, konnte die Gefahr einer Background-Join-Attack gezeigt werden. Aus diesem Grund verwendet das entwickelte $Ft EDPP_2GA$-Protokoll das Prinzip der Fragmentierung zum Schutz der Daten. Durch die strikte Trennung der Schritte „k-Anonymisierung “, „Validierung der t-Closeness“, „Berechnung gemeinsamer Äquivalenzklassen“ sowie „Fragmentierung und Veröffentlichung“ konnte die Korrektheit des Protokolls gezeigt werden. Das Hauptaugenmerk lag auf der Umsetzung einer Anonymisierung unter Beachtung der t-Closeness in einem verteilten Szenario. Betrachtungen zur Laufzeit des Protokolls lagen außerhalb des Fokus dieser Arbeit.

Durch die Adaption des DPP_2GA-Protokolls erfuhr der DF-Algorithmus eine erneute Entwicklung in Form des $Ft EDPP_2GA$-Protokolls. Die Verwendung von Konzepten, die dem Lesenden bekannt sind, sowie der Modularisierung der einzelnen Schritte des Verfahrens, ermöglicht eine nachvollziehbare und anschauliche Einführung in das Themengebiet. In diesem Zusammenhang bedarf eine Modellierungsentscheidung besonderer Hervorhebung. In Abschnitt 5.3.4.1 wurde die Notwendigkeit der Gleichheit der Äquivalenzklassen besprochen. Diese Festlegung dürfte eine erhebliche Einschränkung der Utility bezüglich der Assoziation von Quasi-Identifikatoren (QIDs) und sensiblen Attributen (SAs) bedeuten. Wenngleich im Rahmen dieser Arbeit keine statistische Evaluation des entworfenen Protokolls vorgenommen wurde, so ist davon auszugehen, dass die Forderung einer Gleichheit der Äquivalenzklassen jeder beteiligten Partei in vielen Fällen zu einer maximalen Generalisierung führen dürfte. Durch die Fragmentierung hätte dies nicht den Verlust der Attributwerte der QIDs zur Folge. Der zu erwartende Umfang der Äquivalenzklassen würde die Aussagekraft der QIDs in Bezug auf die SAs schmälern. Es sind Erweiterungen hinsichtlich der Lösung dieses Problems denkbar. Wir werden diese in Abschnitt 6.3 genauer betrachten.

Als Nachteil der Fragmentierung muss der Verlust der statistischen Abhängigkeiten zwischen den sensiblen Attributen herausgestellt werden. Weiterhin handelt es sich bei dem verwendeten Protokoll von Jiang et al. formell nicht um ein Protokoll der Secure-Multiparty-Computation (SMC). Dies wurde in Abschnitt 5.2.2 dargelegt. Erweiterungen, welche diesen Umstand ausräumen, sind bekannt und werden in Abschnitt 6.2.1 besprochen.

Durch die Beseitigung der angesprochenen Probleme steigt die Komplexität des Verfahrens. Dies hätte der Vermittlung des Konzepts an den Lesenden geschadet. Der Fokus der Arbeit lag auf der Evaluierung der Umsetzbarkeit eines verteilten Verfahrens unter Berücksichtigung der t-Closeness. Die Optimalität hinsichtlich der Utility der Daten wurde in den Hintergrund gestellt. Der Autor dieser Arbeit gab daher dem leichter verständlichen Konzept den Vorzug.

6.2 VERWANDTE ARBEITEN

Das Gebiet der Anonymisierung von lokalen und dezentralen Daten ist aufgrund seiner Aktualität ein sehr aktives Forschungsfeld. Um einen Ausblick auf zukünftige Entwicklungen wagen zu können, werden wir uns in Abschnitt 6.2.1 mit aktuellen Erkenntnissen des Gebiets des PPDP befassen.

In Abschnitt 6.2.2 werden wir dezentrale Konzepte der Datenanonymisierung unter Verwendung der k-Anonymisierung und ℓ-Diversity nennen.

6.2.1 *Weitere Konzepte des Privacy-Preserving Data Publishing*

Es existieren Verfahren, welche versuchen den durch die t-Closeness induzierten Utility-Verlust zu minimieren, ohne die Vorteile des Konzepts zu verlieren.

Li et al. selbst erweiterten das Konzept der t-Closeness zu einer weniger restriktiven Ausprägung namens (n, t)-Closeness. In dieser Ausprägung wird die t-Closeness nicht mit Bezug auf die gesamte Veröffentlichung berechnet, sondern bezogen auf eine höhere Generalisierung der betrachteten Äquivalenzklasse, welche mindestens n Tupel enthält [LLV10].

Andere Arbeiten konzentrieren sich auf die der t-Closeness zu Grunde liegenden Metrik, die EMD, um eine Verbesserung der Utility zu erreichen. So bemüht die Arbeit von Sha et al. die Kullback-Leibler-Distanz (KL-Distanz) als Distanzmaß [SLZ10]. Diese ist uns bereits in Kapitel 4 begegnet.

Einen weiteren interessanten Ansatz stellt das Prinzip der *Semantic Privacy* von Brickell et al. dar [BS08, S. 72]. Das Verfahren zur Umsetzung dieses Prinzips ist als *δ-disclosure-privacy* bekannt. Es ermittelt für jede Äquivalenzklasse das logarithmier-

te Verhältnis der Häufigkeit jedes sensiblen Attributwerts der Äquivalenzklasse, zu dessen Häufigkeit in der gesamten Tabelle. Durch die reine Betrachtung des Verhältnisses, das die relativen Häufigkeiten der Attributwerte zueinander haben, entfällt eine Unterscheidung nach quantitativen oder kategorialen Attributwerten. Im Unterschied zu den Metriken der t-Closeness, berücksichtigt die δ-disclosure-privacy nach Brickell et al. auch selten auftretende Attributwerte [BS08, S. 72]. Nach Brickell et al. ist dieses Verfahren auch für den Fall der MSA geeignet [BS08, S. 71]. Die Autoren machen diesbezüglich jedoch keine Aussage zu dem Curse of Dimensionality.

Eine nach Frikken et al. orthogonale, wenngleich wichtige Entwicklung, stellt die Behandlung von wiederkehrenden Veröffentlichungen von Mikrodaten dar [FZ08]. Diese sogenannte *Redistribution* fand bereits unter dem Stichwort *m-Invariance* für die Konzepte k-Anonymisierung und ℓ-Diversity Beachtung [XT07]. Eine Adaption der t-Closeness an dieses Szenario findet sich in Form des SABRE-Frameworks in [Cao+11].

6.2.2 *Weitere Konzepte der Anonymisierung von vertikal partitionierten Daten*

Speziell im medizinischen Umfeld finden sich Arbeiten, welche die Zusammenführung und Anonymisierung von verteilten Daten thematisieren. Diese konzentrieren sich vielfach auf die Verwendung einer Trusted-Third-Party (TTP) [El +11; Kan+08; Kan+09; KJM08].

Die konzeptionell neue Betrachtung von Methoden des Datenschutzes durch die Differential Privacy, förderte das Interesse vieler Autoren an diesem Themengebiet.

Die Betrachtung von Anonymisierungskonzepten des PPDP, in einer verteilten Umgebung ohne die Verwendung einer TTP, stellt derzeit eine Nische dar. Demgemäß ist die Anzahl direkt vergleichbarer Arbeiten überschaubar.

Nachfolgend wollen wir zunächst Konzepte zur Berechnung einer k-Anonymisierung von partitionierten Daten betrachten. Anschließend widmen wir uns Ansätzen, welche die Umsetzung der ℓ-Diversity auf verteilten Daten anstreben.

Da die dezentrale Anonymisierung von horizontal verteilten Daten nicht im Fokus dieser Arbeit steht, wollen wir uns an dieser Stelle nicht näher mit Arbeiten zu diesem Thema befassen. Dem interessierten Lesenden sei an dieser Stelle die folgende Literatur empfohlen: [Zho09; JX09; DTK11].

6.2.2.1 *Konzepte der dezentralen k-Anonymisierung*

In Kapitel 5 konnte gezeigt werden, dass das DPP_2GA-Protokoll Schwächen hinsichtlich des Vergleichs der Äquivalenzklassen sowie der Vereinigung der Teiltabellen aufweist. Diese entstanden durch die vertikale Partitionierung der Tabelle.

Die Schwachstelle, bezüglich des Vergleichs der Äquivalenzklassen, begründet sich in der Tatsache, dass es sich bei DPP_2GA nicht um ein SMC-Protokoll handelt [JC05, S. 168].

Jiang et al. griffen die Kritik an dem eigenen Protokoll auf und entwickelten eine Erweiterung, welche die genannten Mängel aufhebt: Das *DkA*-Protokoll [JC06]. In ihrer Arbeit entwickelten Jiang et al. zu diesem Zweck ein Verfahren zur Secure Set Intersection (SSI). Dieses vermittelt den an dem Verfahren beteiligten Parteien ausschließlich das Wissen darüber, ob die Vereinigung beider Teiltabellen eine global k-anonyme Tabelle erzeugen würde oder nicht. Im Gegensatz zu dem DPP_2GA-Protokoll ist für die beteiligten Parteien nicht nachvollziehbar welchen Umfang der Schnitt der Teilmengen besitzt. Ein weiterer Vorteil der Verwendung einer SSI besteht in der Möglichkeit des erneuten Vergleichs einer Äquivalenzklasse, die in der nächsten Iterationsstufe unverändert ist. Ohne die Verwendung eines SSI-Mechanismus könnte dies die Sicherheit des Verfahrens gefährden [JC05, S. 176].

Eine sichere Vereinigung der global k-anonymen Teiltabellen vollzieht das *DkA*-Protokoll von Jiang et al. durch die Verwendung kommutativer Verschlüsselungsfunktionen [JC06, S. 317].

Wenngleich dieses Protokoll eine erhebliche Erweiterung der Sicherheit dezentraler Anonymisierung von Daten darstellt, so erlangt es diese auf Kosten der Einfachheit des Verfahrens.

Ein weiterer Ansatz zur dezentralen Berechnung einer k-Anonymisierung findet sich in Mohammed et al. [MFD11]. Dieser sieht eine Top-Down-Strategie zur Anonymisierung verteilter Daten vor. Das Verfahren beschränkt sich nicht auf die Verwendung von ausschließlich zwei Parteien. Initial sendet jede an dem Verfahren beteiligte Partei eine komplett generalisierte Version ihrer Teiltabellen an sämtliche anderen Parteien. Diese können die Tabellen gefahrlos vereinen und besitzen nun lokal eine maximal anonymisierte Tabelle. In jedem weiteren Schritt des Protokolls prüft jede Partei, ob eine weitere Spezialisierung der von ihr beigetragenen Attribute eine Verletzung der Privatsphäre darstellen würde. Ist dies nicht der Fall, einigen sich die Parteien auf die beste Spezialisierung. Eine besondere Eigenschaft dieses Protokolls ist seine Robustheit gegenüber einem böswilligen Angreifer.

Dieses Verfahren bildet aufgrund des Paradigmas der Top-Down-Strategie eine fundamental andere Herangehensweise als das Bottom-Up-Verfahren von Jiang et al. [JC06]. Einige Autoren vertreten die Auffassung, dass eine Top-Down-Strategie in einem verteilten Szenario effektiver sei als eine Bottom-Up-Strategie [JX09, S. 197].

6.2.2.2 *Konzepte der dezentralen ℓ-Diversity*

Das zuvor erwähnte Top-Down-Protokoll von Mohammed et al. ist nach Aussage des Autors auf das Konzept der ℓ-Diversity erweiterbar [MFD11, S. 584].

Einen ähnlichen Ansatz wählten Fung et al. [Fun+11]. Sie entwickelten eine Generalisierung der Konzepte k-Anonymisierung und ℓ-Diversity: Die *LKC*-Privacy [Fun+11, S. 5]. Diese erlaubt eine weniger restriktive Gestaltung der Äquivalenzklassen. Ein Vergleich der *LKC*-Metrik zu den Metriken der k-Anonymisierung und ℓ-Diversity bezüglich des Schutzes der Daten erfolgte bislang nicht.

Das Protokoll von Fung et al. besticht jedoch durch eine sehr schnelle Berechnung der Anonymisierung. Nach Angaben der Autoren ist die Anonymisierung mittels *LKC*-Privacy in einem verteilten Szenario innerhalb weniger Minuten möglich [Fun+11, S. 13f]. Die Anonymisierung einer Tabelle mithilfe des *DkA*-Protokolls erfordert annähernd 10 Tage [JC06, S. 331]. In beiden Fällen wurde das *Adults census dataset* verwendet [KB98]. Dieses besteht aus 14 Attributen, welche durch annähernd 50 000 Tupel abgebildet werden.

Eine Interessante Alternative zu den vorher genannten Ansätzen bietet der *Sequential Algorithm* von Goldberger et al. [GT10]. Dieser Algorithmus geniert initial zufällige Äquivalenzklassen. Die QID der Tupel, welche in diesen als *Cluster* bezeichneten Äquivalenzklassen vorliegen, werden durch Generalisierung vereinheitlicht. Ausgehend von dem, durch eine Utility-Metrik berechneten, minimalen Informationsverlust verschiebt der Algorithmus einzelne Tupel zwischen den Clustern [GT10, S. 163]. Tassa et al. übertrugen diesen Algorithmus auf Szenarien der horizontalen und vertikalen Partitionierung [TG12].

6.3 AUSBLICK

In den vorherigen Abschnitten wurden die Ergebnisse der Diplomarbeit rekapituliert sowie weitere Ansätze aus verwandten Themengebieten betrachtet.

An dieser Stelle wollen wir das zuvor Gesagte aufgreifen und analysieren wie die vorliegende Arbeit weitergeführt werden kann.

Ansatzpunkte für die Weiterentwicklung der vorliegenden Arbeit sind in vielerlei Hinsicht denkbar. Diese lassen sich in verschiedene Arbeitsbereiche gliedern. Ein Ansatzpunkt wäre die Erweiterung der Ground-Distance-Metriken, welche von der t-Closeness verwendet werden. Eine weitere Möglichkeit bestünde in der Verwendung eines andern Anonymisierungskonzepts als der t-Closeness. Letztendlich bietet die Verbesserung kryptographischer Protokolle ein weiteres Forschungsfeld.

Widmen wir uns dem ersten Punkt. Die Ground-Distance-Metriken der EMD stellen eine besondere Herausforderung dar. Sowohl die Ordered Distance als auch die

Hierarchical Distance sind nicht für die Verwendung mit mehreren sensiblen Attributen ausgelegt. Im Zuge dieser Diplomarbeit konnte eine Veröffentlichung mehrerer sensibler Attribute durch Fragmentierung der Daten erreicht werden. Eine Erweiterung der genannten Distanzmaße, hinsichtlich der Betrachtung mehrerer sensibler Attribute und deren statistischer Abhängigkeit, böte den Vorteil des Erhalts der Assoziation verbundener Attributwerte. In Anhang A wurde bereits ein erster Ansatz in diese Richtung skizziert. Dieser berücksichtigt jedoch nicht die statistische Abhängigkeit der sensiblen Attribute.

Ein direkt verbundenes Forschungsgebiet stellt die Verwendung anderer Konzepte des Datenschutzes in einer verteilten Umgebung dar. In den vorhergehenden Abschnitten wurden Verfahren benannt, welche auf den Prinzipien der t-Closeness aufbauen. Diese verwenden unterschiedliche Distanzmaße. Insbesondere die Verwendung der Entropie-Maße stellt eine interessante Möglichkeit der Betrachtung statistischer Abhängigkeiten dar. Durch die Fokussierung auf den Informationsgehalt der Attributwerte wird das Problem der Unterscheidung des Skalenniveaus umgangen. Eine weitere lohnende Betrachtung wäre die Adaption von Konzepten, deren Schutz auf Fragmentierung der Daten beruht. Das et al. entwickelte ein derartiges Konzept namens `Decomposition+` zum Schutz der Daten mittels ℓ-Diversity. Eine Erweiterung des Verfahrens auf das Konzept der (n, t)-Closeness wäre laut Das et al. denkbar [DB12, S. 10].

Die Betrachtung statistischer Abhängigkeiten in einer verteilten Umgebung konfrontiert uns mit dem Problem der Background-Join-Attack. Durch sie ergibt sich ein Paradigmenwechsel der SAs von informativen Attributen hin zu identifizierenden Attributen.

Unter der Voraussetzung, dass die SAs einer Tabelle nicht-identifizierend wirken, wären Protokolle denkbar, welche mithilfe der SMC eine Statistik über die SAs erstellen. Eine derartige Statistik könnte, im Schritt der lokalen Berechnung der Anonymisierung, zur Validierung der t-Closeness verwendet werden. Der weitaus effizientere Ansatz wäre jedoch die Verwendung von Top-Down-Algorithmen zur Berechnung der Anonymisierung. An dieser Stelle wären Vorgehensweisen wie in den Arbeiten von Fung et al. [Fun+11] und Mohammed et al. [MFD11] denkbar. Das Verfahren von Jiang et al. [JC06] bietet die Möglichkeit der Verwendung von Top-Down-Algorithmen. Jedoch bleiben die konkreten Werte der sensiblen Attribute durch die Verwendung des SSI-Protokolls verborgen. Sie können daher nicht für die Berechnung einer t-Closeness herangezogen werden.

Die Diplomarbeit konnte zeigen, dass die Größe der Äquivalenzklassen eine unmittelbare Gefahr für eine globale t-Closeness darstellt. Dieses Problem besteht bei Bottom-Up-Mechanismen wie dem in dieser Arbeit verwendetem DPP_2GA-Protokoll. Aus diesem Grund wurde, für das in dieser Arbeit vorgeschlagenen Protokoll, eine

Gleichheit der Äquivalenzklassen statt einer Äquivalenz gefordert. Eine Erweiterung dieses Protokolls sollte sich darauf konzentrieren, dass durch einen Schnitt der Umfang der initialen Äquivalenzklassen nicht unterschritten wird. Eine Vergrößerung der Äquivalenzklasse stellt hingegen kein Problem dar. Wie in der Arbeit gezeigt wurde, ist die Vereinigung zweier Äquivalenzklassen, welche der t-Closeness genügen, stets t-close. Bei der Verwendung eines Protokolls ohne die Eigenschaften der SMC müsste zudem bewiesen werden, dass keine Rückschlüsse auf private Daten möglich sind, die durch die Festlegung unterschiedlicher Größen für Äquivalenzklassen entstehen.

Abschließend soll die Notwendigkeit der Verbesserung von SMC-Protokollen herausgestellt werden. Die Kosten bezüglich Rechenzeit und Kommunikationsaufwand derartiger Protokolle unterscheiden sich, je nach Funktionalität des verwendeten Protokolls, teilweise drastisch [TG12, 11:36ff]. Es existieren Abhandlungen, welche den Aspekt der zu erwartenden Kosten eines SMC-Protokolls dem Mehrgewinn gegenüberstellen, der durch die Ausführung des Protokolls zu erwarten ist [Ner+11].

Ein weiteres interessantes Thema ergibt sich aus der Erweiterung des in dieser Arbeit vorgeschlagenen Protokolls, auf mehr als zwei Parteien. Jiang et al. erwägen zur Lösung dieses Problems die iterative, paarweise Ausführung des Protokolls [JC06, S. 332]. Eine Erhöhung der Teilnehmer stellt jedoch stärkere Anforderungen an das zuvor genannte Problem der Größe der Äquivalenzklassen. Die Verwendung eines Protokolls, welches nach dem Top-Down-Prinzip vorgeht, erscheint in dieser Hinsicht lohnenswert. Es existieren bereits Ansätze, welche die Verwendung von mehr als zwei Parteien vorsehen [Fun+11; MFD11; TG12].

Aufgrund der genannten Möglichkeiten bleibt der Schutz der Privatsphäre ein dynamisches Forschungsfeld. Die subjektive Wahrnehmung des Begriffs bietet großen Spielraum für die Entwicklung weiterführender Konzepte des Datenschutzes. Die Betrachtung mehrerer sensibler Attribute sowie die Möglichkeit eines verteilten Szenarios sollten substanzielle Eigenschaften zukünftiger Konzepte bilden.

A GROUND DISTANCE FÜR MEHRERE SENSIBLE ATTRIBUTE

Wie in Kapitel 4 gezeigt ist die Unterteilung der Attribute in qualitative und quantitative Attributwerte geboten. Wir wollen daher auch hier nach dieser Vorgehensweise verfahren und beschäftigen uns zunächst mit der Ground-Distance zweier quantitativer Attribute.

A.1 VERWENDUNG DES SKALENNIVEAUS

In Abschnitt 2.1.2 haben wir gesehen, dass gewisse Merkmalsausprägungen einer Rangfolge unterliegen. Dies ist bei Merkmalen der Fall, welche mindestens dem ordinalen Skalenniveau angehören. Um die Ausprägungen dieser Werte geeignet referenzieren zu können, wollen wir den Begriff des *Rangwerts* einführen. Es sei vorausgesetzt, dass keine *Bindungen* zwischen den Untersuchungseinheiten existieren, d. h. jeder Merkmalsträger besitzt eine individuelle Merkmalsausprägung [Haro9, S. 139]. Wir definieren nach [CKo8, S. 18]:

Definition 36 (Rangwertreihe)**.** *Seien* $\mathcal{A}_X = \{x_1, \ldots, x_m\}$ *die Beobachtungswerte eines metrisch skalierten Merkmals, so wird die aufsteigend geordnete Auflistung der Beobachtungswerte*

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(j)} \leq \ldots \leq x_{(m-1)} \leq x_{(m)}$$

als Rangwertreihe *bezeichnet.*

Dabei repräsentiert $x_{(j)}$ die Position eines $x_i \in \mathcal{A}_X$ an der j-ten Stelle der Rangwertreihe. $x_{(j)}$ ist somit der *Rang von* x_i und wird zumeist als *Rangwert* oder *Rangzahl* bezeichnet. Wir wollen in diesem Zusammenhang eine Funktion $R : \mathcal{A}_X \longrightarrow \mathbb{N}$ definieren, welche einem Beobachtungswert x_j eine natürliche Zahl – seinen Rang – zuordnet:

$$R(x_i) = x_{(j)} = j \tag{A.1}$$

In der Literatur zu statistischen Verfahren wird der Rangwert u. a. zur Ermittlung von Lage- sowie Zusammenhangsmaßen [Koho5; CKo8; Haro9] verwendet. In diesem Kontext ist es sinnvoll, Bindungen zu gewichten. Für den weiteren Verlauf dieser Arbeit ist es jedoch unwesentlich, ob ein Merkmalswert mehr als einmal in der

Stichprobe vorhanden ist. Wir wollen in den nachfolgenden Kapiteln davon ausgehen, dass jeder Merkmalsträger eine einzigartige Ausprägung des Merkmals besitzt und dadurch seine Position innerhalb der Rangwertreihe eindeutig bestimmbar ist[1]. Wir bezeichnen den ersten Rangwert $x_{(1)}$ als *Minimum* und den letzten Rangwert $x_{(max)}$ als *Maximum* der Menge $\mathcal{A}_X$.

Der Bereich, in dem die Werte der Beobachtungswerte liegen, wird als *Spannweite* oder *Range* bezeichnet. Er errechnet sich aus der Differenz des Maximums und des Minimums der Rangwertreihe und wird im Allgemeinen als

$$R = x_{(max)} - x_{(1)} \tag{A.2}$$

notiert.

Nach Kohn verfügen die Rangzahlen über eine nützliche Eigenschaft: „[Sie] besitzen einen metrischen Abstand, der stets eins ist." [Koh05, S. 121]. Wir wollen diesen Umstand festhalten:

$$\forall i \in \mathbb{N} : |x_{(i)} - x_{(i+1)}| = 1 \tag{A.3}$$

Durch das Überführen der Merkmalsausprägungen in ihre Rangwerte wurde ihre ursprüngliche Metrik aufgehoben. In Definition 36 wurde somit ein metrisch skaliertes Merkmal in ein ordinal skaliertes Merkmal überführt. Dieser Vorgang wird als *Skalentransformation* bezeichnet. Nach Kohn versteht man unter einer Skalentransformation „ [. . .] die Übertragung der Skalenwerte in Werte einer anderen Skala, wobei die Ordnungseigenschaften der Skala erhalten bleiben müssen." [Koh05, S. 15].

In diesem Zusammenhang sei auf die wichtige Feststellung verwiesen, dass eine Skalentransformation nur von einem höheren auf ein niedrigeres Skalenniveau erfolgen kann.

A.2 GROUND-DISTANCE FÜR ZWEI QUANTITATIVE ATTRIBUTWERTE

Wir wollen das zuvor erlangten Kenntnisse nun nutzen um eine Ground-Distance für mehr als ein sensibles Attribut zu definieren. Nach bestem Wissen und Gewissen des Verfassers existiert bislang keine Publikation, welche eine Ground-Distance für mehr als ein sensibles Attribut definiert.

Um uns einer solchen Kostenfunktion zu nähern betrachten wir erneut die Ordered Distance aus Gleichung (4.8). Sie definiert den Abstand zweier Elemente einer

1 Sollte dem nicht so sein, so kann eine eindeutige Zuordnung durch eine surjektive Abbildung der Beobachtungswerte auf die Rangwertreihe erfolgen.

geordneten Menge als die Anzahl der zwischen ihnen liegenden Elemente. Das Resultat wird auf das Intervall $[0, 1]$ skaliert, um das Ergebnis der Berechnung der EMD ebenfalls auf dieses Intervall zu beschränken.

Durch die Hinzunahme eines weiteren Attributs verlassen wir den eindimensionalen Raum der bisherigen Ereignisse und müssen diese Metrik in den sich ergebenen zweidimensionalen Raum übertragen. Wir wollen an dieser Stelle erneut davon ausgehen, dass die beiden sensiblen Attribute mit X_1 und X_2 bezeichnet seien. Ihre Ausprägungen seien demnach $\mathcal{A}_{X_1} = \{x_{1i} : 1 \leq i \leq m\}$ sowie $\mathcal{A}_{X_2} = \{x_{2j} : 1 \leq j \leq l\}$ mit den zugehörigen Häufigkeitsverteilungen $X_1 = \{f(x_{11}), \ldots, f(x_{1m})\}$ und $X_2 = \{f(x_{21}), \ldots, f(x_{2l})\}$ respektive. Wie in der Ordered Distance von Li et al. wollen wir die Anzahl der Elemente zwischen zwei Elementen als Maß für deren Abstand betrachten.

Eine Metrik, welche das Gewünschte leistet und uns namentlich bereits aus Abschnitt 4.3.2.1 geläufig ist, ist die MH-Distanz. Diese Metrik spiegelt im mehrdimensionalen Raum die Distanz zweier Punkte wider:

$$d_{MH}(\omega_i \| \omega_j) = \sum_{k=1}^{m=2} |x_{ik} - x_{jk}| \tag{A.4}$$

Zu beachten ist, dass der Laufindex durch 2 begrenzt ist, da wir uns auf den zweidimensionalen Raum beschränken.

Eine Einfache Adaption dieser Metrik würde uns jedoch Abstandsmaße unterschiedlicher Größe, abhängig von den verwendeten Attributen und deren Werten liefern. Um ein einheitliches Ergebnis zu erhalten müssen wir daher die Attributwerte einer Skalentransformation unterziehen. Zur Berechnung der MH-Distanz verwenden wir daher die resultierenden Rangzahlen (vgl. Definition A.1).

$$d_{MH}(\omega_i \| \omega_j) = \sum_{k=1}^{m=2} |R(x_{ik}) - R(x_{jk})| \tag{A.5}$$

Die MH-Distanz ergibt sich folgerichtig aus dem Absolutbetrag der Differenz der Ränge aller Attributwerte der betrachteten Tupel.

Aus Abschnitt 4.3.2.3 ist uns bereits bekannt, dass wir auch das Ergebnis einer mehrdimensionalen Ordered Distance auf das Einheitsintervall normieren müssen. Dieses Verfahren wenden wir nun ebenfalls auf die MH-Distanz der Ränge an. Zu diesem Zweck betrachten wir die durch die Metrik erreichbaren Extremwerte, d. h. das Maximum sowie das Minimum und dividieren das

In Analogie zur Ordered Distance bezeichnen wir die Kostenfunktion $c(\cdot,\cdot)$ der EMD für den MSA-Fall mit *MSA Ordered Distance* und definieren diese wie folgt:

$$c(x_{ik}, x_{jk}) = msa\_ordered\_dist(x_{ik}, x_{jk}) = \frac{\sum_{k=1}^{m=2} |R(x_{ik}) - R(x_{jk})|}{\sum_{k=1}^{m=2} R(x_{(max)k}) - R(x_{1k})} \tag{A.6}$$

Wir wollen beispielhaft den Abstand der Tupel 1 und 5, über den quantitativen Attributen `Gehalt` und `L-WBC` aus Tabelle 5.2 berechnen.

Beispiel 17.

$$\begin{aligned}
&msa_ordered_dist(x_{1k}, x_{5k}) \\
&= \frac{|R(x_{1\texttt{Gehalt}}) - R(x_{5\texttt{Gehalt}})| + |R(x_{1\texttt{L-WBC}}) - R(x_{5\texttt{L-WBC}})|}{(R(x_{(max)\texttt{Gehalt}}) - R(x_{1\texttt{Gehalt}})) + (R(x_{(max)\texttt{L-Wbc}}) - R(x_{1\texttt{L-Wbc}}))} \\
&= \frac{|R(3000) - R(11000)| + |R(11000) - R(3000)|}{(R(x_{(max)\texttt{Gehalt}}) - R(x_{1\texttt{Gehalt}})) + (R(x_{(max)\texttt{L-Wbc}}) - R(x_{1\texttt{L-Wbc}}))} \\
&= \frac{|1-9| + |9-1|}{8+8} = \frac{8+8}{16} = 1
\end{aligned}$$

Der Abstand ist folglich maximal. Diese Beobachtung leuchtet ein, sind doch die attributweise betrachteten Abstände ebenfalls maximal.

Die nachfolgende Grafik stellt die Extrema der *msa_ordered_distance* dar und verdeutlicht ihre Vorgehensweise.

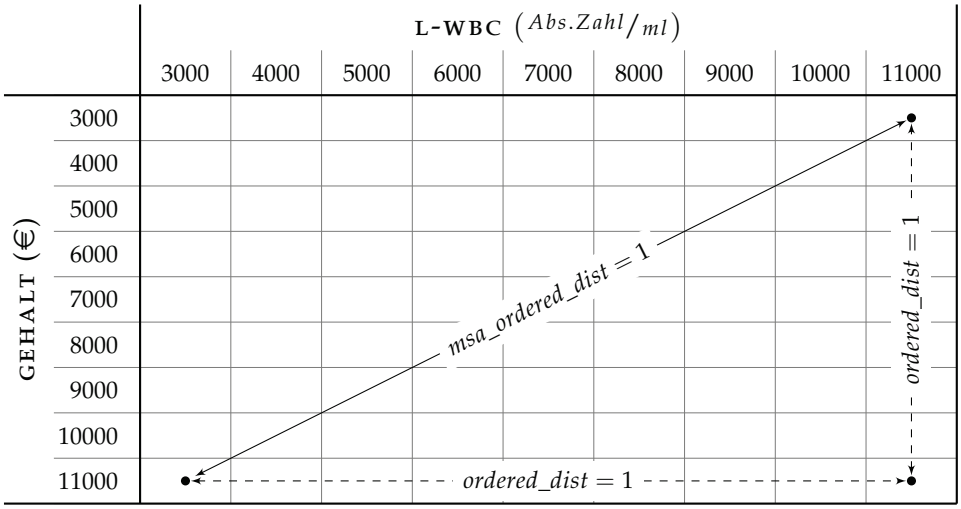


Abbildung A.1: Ordered Distance zweier quantitativer Attribute

Die EMD mithilfe der *msa_ordered_distance* kann nun analog zur EMD für einzelne Attribute unter der *ordered_distance* berechnet werden. Hierzu wird eine Verteilung sukzessive durch das Verschieben von Wahrscheinlichkeitsmasse eines Elements zu seinem Nachbarn transformiert.

Beispiel 18. Sei die Verteilung der Äquivalenzklasse E_1 wie in Abbildung A.2. Wir wollen E_1 mit V vergleichen (mit der Einschränkung auf die quantitativen Attribute `Gehalt` und `L-WBC`). Wir summieren daher sämtliche Verschiebungen an Wahrscheinlichkeitsmasse zu den jeweiligen Nachbarn auf und erhalten das Ergebnis der EMD. Beispielsweise sind die Tupel 1 und 2, also die Attributwertpaare $(3000\,€, 11000\ ^{WBC}/_{ml})$ und $(4000\,€, 10000\ ^{WBC}/_{ml})$ benachbart. Dies ist der Fall da die Attributwerte 3000 € und 4000 € sich nur um einen Rang unterscheiden. Die Differenz beträgt:

$$
\begin{aligned}
&msa\_ordered\_dist(x_{1k}, x_{2k}) \\
&= \frac{|R(x_{1\texttt{Gehalt}}) - R(x_{2\texttt{Gehalt}})| + |R(x_{1\texttt{L-WBC}}) - R(x_{2\texttt{L-WBC}})|}{(R(x_{(max)\texttt{Gehalt}}) - R(x_{1\texttt{Gehalt}})) + (R(x_{(max)\texttt{L-Wbc}}) - R(x_{1\texttt{L-Wbc}}))} \\
&= \frac{|R(3000\,€) - R(4000\,€)| + |R(11000\ ^{WBC}/_{ml}) - R(10000\ ^{WBC}/_{ml})|}{(R(11000\,€) - R(3000\,€)) + (R(x_{(max)\texttt{L-Wbc}}) - R(x_{1\texttt{L-Wbc}}))} \\
&= \frac{1+1}{8+8} = \frac{1}{8}
\end{aligned}
$$

In diesem Beispiel gilt dies für alle Nachbarn. Daher betragen die Kosten zum Transport von Masse jeweils $1/8$. Es muss folglich nur noch die Summe der zu transportierenden Wahrscheinlichkeitsmasse betrachtet werden, welche sich analog zur Ordered Distance von Li et al. berechnet.

Wir wollen an dieser Stelle nicht auf Details der Kostenfunktion eingehen oder ihre Eigenschaften konkret Beweisen. Dies liegt außerhalb des Fokus dieser Arbeit. Vielmehr soll an dieser Stelle eine Intuition einer generalisierten Kostenfunktion gegeben werden. Im nächsten Abschnitt werden wir diese Betrachtung für zwei kategoriale Attribute wiederholen.

A.3 GROUND-DISTANCE FÜR ZWEI KATEGORIALE ATTRIBUTWERTE

Eine Adaption der Hierarchical Distance kann nicht ohne weiteres erfolgen. Dies liegt in der Verwendung eines VGH begründet. Ein VGH bildet stets ein subjektives Maß. Die möglichen Generalisierungs-Schritte sind dem Design des VGH ausgeliefert. Für den Fall des Global Recodings erlaubt ein VGH niedriger Höhe naturgemäß nur wenige Generalisierungs-Schritte bis zur Unterdrückung eines Attributs. Wie

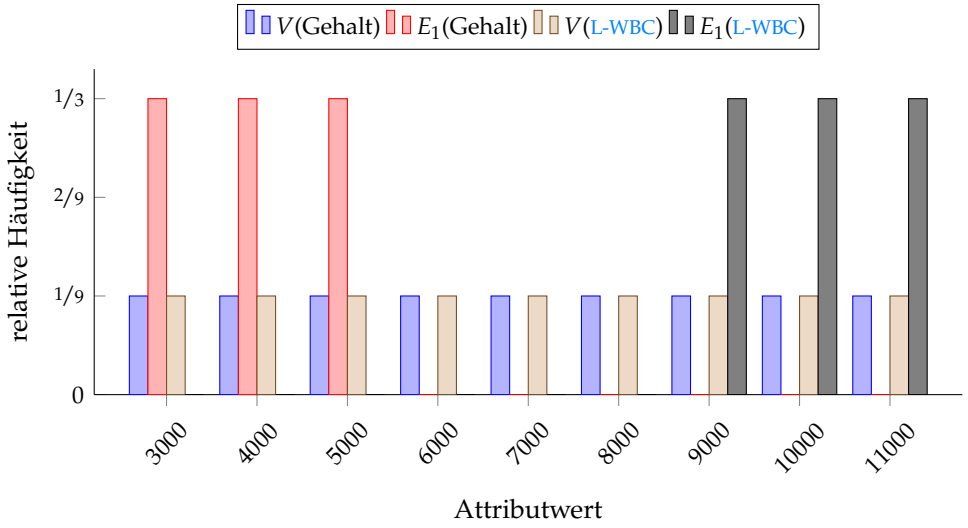


Abbildung A.2: Verteilung nicht korrelierender Attribute

wir in Kapitel 4 gesehen haben findet ein VGH jedoch nicht nur bei Algorithmen der Generalisierung und Unterdrückung von Attributwerten Verwendung. Es wurde ein VGH verwendet um die Nähe zweier kategorialer Attributwerte zu charakterisieren. Wir wollen an einem Beispiel erläutern wie die Höhe des VGH die t-Closeness einer Tabelle beeinflusst.

Beispiel 19. Betrachten wir erneut Tabelle 4.2 mit den Äquivalenzklassen E_1, E_2 und E_3 sowie den bekannten Verteilungen ihrer Ausprägungen aus Tabelle 4.3. Aus Beispiel 12 ist uns bekannt, dass die t-Closeness der Äquivalenzklasse E_2 zur Gesamtveröffentlichung V $^8/_{27}$ in Bezug auf das kategoriale Attribut `Krankheit` beträgt. Die Berechnung der restlichen Abstandsmaße zwischen der Veröffentlichung und den Äquivalenzklassen ist nicht von Interesse. Wir wollen daher vorgreifen und geben diese vor:

$$d_{EMD}(E_1 \| V) = {}^4/_9 \tag{A.7}$$
$$d_{EMD}(E_2 \| V) = {}^8/_{27} \tag{A.8}$$
$$d_{EMD}(E_3 \| V) = {}^6/_{27} \tag{A.9}$$

Betrachten wir den VGH 4.3 genauer, so fallen die der Abbildung grau gekennzeichneten Knoten auf. Diese spiegeln keine Attributwertausprägungen der Veröffentlichung wider. Erstellen wir einen alternativen VGH ohne diese Knoten, so erhalten wir einen Baum wie er in Abbildung A.3 zu sehen ist. Berechnen wir ausgehend von diesem VGH die t-Closeness der Äquivalenzklassen, ergibt sich ein anderes Bild

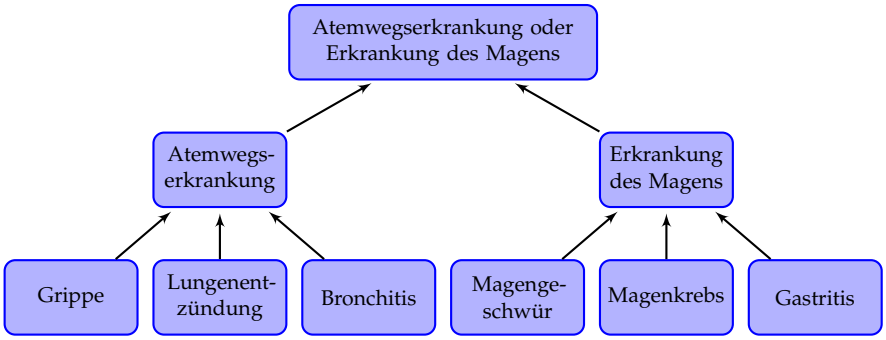


Abbildung A.3: Verkürzter VGH des sensiblen Attributs `Krankheit`

der Nähe zwischen den Attributwerten Erneut wollen wir nicht auf die Details der Berechnung eingehen.

$$d_{EMD}(E_1 \| V) = 4/9 \tag{A.10}$$
$$d_{EMD}(E_2 \| V) = 4/9 \tag{A.11}$$
$$d_{EMD}(E_3 \| V) = 4/9 \tag{A.12}$$

Der Berechnung zur Folge sind sämtliche Äquivalenzklassen gleich „nahe“ an der Verteilung der Veröffentlichung. Dies widerstrebt der Betrachtung der Tabelle 4.2. Die Äquivalenzklassen E_2 und E_3 sollten eine größere Nähe zu V aufweisen, da sie eine größere Vielfalt der Attributwerte beinhalten. Dieses Phänomen liegt in dem Einfluss der Höhe des VGH begründet. Es verdeutlicht gleichzeitig, dass eine Adaption des Ursprünglichen Verfahrens der Hierarchical Distance schwierig ist. Eine Graphentheoretische Betrachtung zur Entwicklung eines neuen Abstandsmaßes steht jedoch außerhalb des Blickfelds dieser Arbeit.

B RECHENBEISPIELE

B.1 KULLBACK-LEIBLER DISTANZ

Die Kullback-Leibler-Distanz (KL-Distanz) ist nach Kullback et al. [KL51] wie folgt definiert:

Definition 37 (Kullback-Leibler-Distanz)**.** *Seien X_E und X_V Häufigkeitsverteilungen, mit den relativen Häufigkeiten $X_E = \{f_E(x_1), \ldots, f_E(x_m)\}$ sowie $X_V = \{f_V(x_1), \ldots, f_V(x_m)\}$ der Ausprägungen $\mathcal{A}_X = \{x_i, 1 \leq i \leq m\}$ eines Merkmals X. Dann gilt:*

$$d_{KL}(X_E \| X_V) := \sum_i f_E(x_i) \cdot \log_2 \left(\frac{f_E(x_i)}{f_V(x_i)} \right) \tag{B.1}$$

Nach Li et al. ist die KL-Distanz für unmögliche Ereignisse, also Ereignisse deren Ausprägungen die Häufigkeit 0 besitzt, nicht definiert [LLV10, S. 950]. Für das unmögliche Ereignis $f(x_i) = 0$ gelte daher für den Summanden $i : f_E(x_i) \log_2 {}^{f_E(x_i)}/_{f_V(x_i)} = 0$. Dies ist eine vernünftige Annahme, wie sie unter anderem von Schmitt getroffen wird. Diese Annahme stellt sicher, dass ein unmögliches Ereignis keinen Einfluss auf die Information eines Merkmals haben kann [Scho6a, S. 186 & 188].

Nachfolgend findet sich ein Beispiel zur Berechnung der KL-Distanz zwischen der Häufigkeitsverteilung der Äquivalenzklassen sowie der Häufigkeitsverteilung der gesamten Tabelle aus Abbildung 4.2. Dieses verdeutlicht, die Vernachlässigung der Semantik der Attributwerte bei der Berechnung des Abstands.

Beispiel 20. Wir betrachten Tabelle 4.2 mit ihren drei Äquivalenzklassen $E_1 = \{1, 2, 3\}$, $E_2 = \{4, 5, 6\}$ sowie $E_3 = \{7, 8, 9\}$. Aus der Tabelle 4.2 können wir die relativen Häufigkeiten des Merkmals (Attributwerts) `Krankheit` ableiten. Diese sind in Tabelle 4.3 für die gesamte Veröffentlichung V sowie die entsprechenden Äquivalenzklassen aufgelistet. Am Beispiel der Äquivalenzklasse E_1 wollen wir uns vergegenwärtigen, dass die KL-Distanz in diesem Falle kein geeignetes Mittel zur Abstandsberechnung der Verteilungen darstellt. Nach Tabelle 4.3 und Gleichung B.1

erhalten wir exemplarisch für den Abstand von der Häufigkeitsverteilung der Äquivalenzklasse E_1 zur Verteilung der gesamten Tabelle V:

$$\begin{aligned}
d_{KL}(X_{E_1} \| X_V) &= \sum_{i=1}^{6} f_{E_1}(x_i) \cdot \log_2 \left(\frac{f_{E_1}(x_i)}{f_V(x_i)} \right) \\
&= f_{E_1}(x_1) \cdot \log_2 \left(\frac{f_{E_1}(x_1)}{f_V(x_1)} \right) + f_{E_1}(x_2) \cdot \log_2 \left(\frac{f_{E_1}(x_2)}{f_V(x_2)} \right) + \\
&\quad + f_{E_1}(x_3) \cdot \log_2 \left(\frac{f_{E_1}(x_3)}{f_V(x_3)} \right) + f_{E_1}(x_4) \cdot \log_2 \left(\frac{f_{E_1}(x_4)}{f_V(x_4)} \right) + \\
&\quad + f_{E_1}(x_5) \cdot \log_2 \left(\frac{f_{E_1}(x_5)}{f_V(x_5)} \right) + f_{E_1}(x_6) \cdot \log_2 \left(\frac{f_{E_1}(x_6)}{f_V(x_6)} \right) \qquad \text{(B.2)} \\
&= 0 \cdot \log_2 \left(\frac{0}{1/9} \right) + 0 \cdot \log_2 \left(\frac{0}{2/9} \right) + 0 \cdot \log_2 \left(\frac{0}{2/9} \right) + \\
&\quad + \frac{1}{3} \cdot \log_2 \left(\frac{1/3}{1/9} \right) + \frac{1}{3} \cdot \log_2 \left(\frac{1/3}{2/9} \right) + \frac{1}{3} \cdot \log_2 \left(\frac{1/3}{2/9} \right) \\
&\approx 0 + 0 + 0 + 0,52832 + 0,19499 + 0,19499 \\
&\approx 0,91830
\end{aligned}$$

Aufgrund der Tatsache, dass die weiteren relativen Häufigkeiten aus Tabelle 4.3 identisch sind, würde die Verwendung der KL-Distanz suggerieren, dass alle Äquivalenzklassen gleich nahe zueinander sind, denn

$$d_{KL}(X_{E_2} \| X_V) \approx 0,91830 \qquad \text{(B.3)}$$

$$d_{KL}(X_{E_3} \| X_V) \approx 0,91830. \qquad \text{(B.4)}$$

Dies widerspricht jedoch unserer, in Beispiel 9 entwickelten Auffassung, dass dem nicht so ist.

LITERATUR

[AES03] Rakesh Agrawal et al. „Information sharing across private databases“. In: *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*. Hrsg. von Alon Y. Halevy et al. San Diego, CA, USA: ACM, 2003, S. 86–97.

[AFN06] Karl Aberer et al., Hrsg. *Proceedings of the 22nd International Conference on Data Engineering*. (Atlanta, Georgia, USA, 3.–7. Apr. 2006). IEEE Computer Society, 2006.

[Agg+05] Gagan Aggarwal et al. „Approximation Algorithms for k-Anonymity“. In: *Journal of Privacy Technology* (Nov. 2005). Das Journal wurde eingestellt. Eine Kopie findet sich unter der URL, S. 246–258. URL: http://ilpubs.stanford.edu:8090/645/.

[Agg05] Charu C. Aggarwal. „On k-anonymity and the curse of dimensionality“. In: *Proceedings of the 31st International Conference on Very Large Data Bases*. (Trondheim, Norwegen). Hrsg. von Klemens Böhm et al. ACM, 2005, S. 901–909.

[Agr+02] Rakesh Agrawal et al. „Hippocratic databases“. In: *Proceedings of 28th International Conference on Very Large Data Bases,* (Hong Kong, China, 20.–23. Aug. 2006). Morgan & Kaufmann Publishers, 2002, S. 143–154.

[AHK01] Charu C. Aggarwal et al. „On the Surprising Behavior of Distance Metrics in High Dimensional Spaces“. In: *Database Theory - ICDT 2001, 8th International Conference*. (London, England, 4.–6. Jan. 2001). Hrsg. von Jan Van den Bussche et al. Bd. 1973. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2001, S. 420–434.

[AJ97] Wendy Alvey et al., Hrsg. *Record Linkage Techniques: 1997 Proceedings of an International Workshop and Exposition*. (Arlington, VA, USA, 20.–21. März 1997). Federal Committee on Statistical Methodology, 1997.

[AYE08] Charu C. Aggarwal et al., Hrsg. *Privacy-Preserving Data Mining: Models and Algorithms*. Bd. 34. Advances in Database Systems. Springer US, 2008.

[Bel03] Richard Ernest Bellman. *Dynamic programming*. Nachdruck der Originalveröffentlichung in 6. Aufl. von 1972: Princeton University Press. Mineola, N.Y., USA: Dover Publ., 2003.

[Bez10] Michele Bezzi. „An information theoretic approach for privacy metrics“. In: *Transactions on Data Privacy* 3.3 (2010), S. 199–215.

[Bro+01] I. N. Bronstein et al. *Taschenbuch der Mathematik*. 5. Aufl. Verlah Harri Deutsch, 2001.

[BS08] Justin Brickell et al. „The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing“. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Hrsg. von Ying Li et al. ACM, 2008, S. 70–78.

[Buc10] Erik Buchmann. „Datenschutz und Data Mining“. Vortragsfolien. Universität Karlsruhe, 2010. URL: http://www.ipd.uni-karlsruhe.de/~buchmann/10SS-Datenschutz/ (besucht am 08. 03. 2013).

[Cao+11] Jianneng Cao et al. „SABRE: a Sensitive Attribute Bucketization and REdistribution framework for *t*-closeness“. In: *The VLDB Journal* (2011).

[Che+09] Bee-Chung Chen et al. „Privacy-Preserving Data Publishing“. In: *Foundations and Trends in Databases* 2.1-2 (2009), S. 1–167.

[Chi+07] Rada Chirkova et al., Hrsg. *Proceedings of the 23rd International Conference on Data Engineering*. (Istanbul, Türkei, 15.–20. Apr. 2007). IEEE Computer Society, 2007.

[Cir+07] V. Ciriani et al. „Microdata protection“. In: *Secure Data Management in Decentralized Systems*. Hrsg. von Ting Yu et al. Bd. 33. Advances in Information Security. New York, NY, USA: Springer Science+Business Media, LLC, 2007, S. 291–321.

[CK08] Erhard Cramer et al. *Grundlagen der Wahrscheinlichkeitsrechnung und Statistik. Ein Skript für Studierende der Informatik, der Ingenieur- und Wirtschaftswissenschaften*. Bd. 2. Springer-Lehrbuch. Springer Berlin / Heidelberg, 2008.

[CT13] Christopher W. Clifton et al. „On syntactic anonymity and differential privacy“. In: *ICDE Workshop on Privacy-Preserving Data Publication and Analysis*. (Brisbane, Australien). Apr. 2013. URL: http://www.openu.ac.il/Personal_sites/tamirtassa/ (besucht am 21. 01. 2013). Eingereicht.

[Dal77] T. Dalenius. „Towards a methodology for statistical disclosure control“. In: *Statistisk Tidskrift* 15.15 (1977), S. 429–444.

[Dal86] Tore Dalenius. „Finding a needle in a haystack - or identifying anonymous census records“. In: *Journal of Official Statistics* 2.3 (1986), S. 329–336.

[DB12] Devayon Das et al. „Decomposition+: Improving ℓ-Diversity for Multiple Sensitive Attributes“. In: *Advances in Computer Science and Information Technology. Computer Science and Engineering*. Hrsg. von Natarajan Meghanathan et al. Bd. 85. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Springer Berlin Heidelberg, 2012, S. 403–412.

[De +10] Sabrina De Capitani di Vimercati et al. „Fragments and Loose Associations: Respecting Privacy in Data Publishing“. In: *The Proceedings of the VLDB Endowment* 3.1 (Sep. 2010).

[De +11] Sabrina De Capitani di Vimercati et al. „Protecting Privacy in Data Release“. In: *Foundations of Security Analysis and Design VI: FOSAD Tutorial Lectures*. Hrsg. von Alessandro Aldini et al. Bd. 6858. Lecture Notes in Computer Science 6. Heidelberg, Deutschland: Springer Berlin / Heidelberg, 2011, S. 1–34.

[Deu07] Bundesrepublik Deutschland. *Gesetz über die Statistik für Bundeszwecke*. Sep. 2007. URL: http://www.gesetze-im-internet.de/bstatg_1987/BJNR004620987.html.

[Deu09] Bundesrepublik Deutschland. *Bundesdatenschutzgesetz*. Aug. 2009. URL: http://dejure.org/gesetze/bdsg.

[Die10] Reinhard Diestel. *Graphentherorie*. Bd. 4. Springer Berlin / Heidelberg, 2010.

[DTK11] K. Doka et al. „KANIS: Preserving k-Anonymity Over Distributed Data“. Proceedings of the 5th International Workshop on Personalized Access, Profile Management, and Context Awareness in Databases. Sep. 2011.

[Dwo11] Cynthia Dwork. „A Firm Foundation for Private Data Analysis“. In: *Communications of the ACM* (Jan. 2011).

[El +11] Khaled El Emam et al. „A secure protocol for protecting the identity of providers when disclosing data for disease surveillance“. In: *Journal of the American Medical Informatics Association* 18.3 (Mai 2011), S. 212–217.

[Fah+07] Ludwig Fahrmeir et al. *Statistik. Der Weg Zur Datenanalyse*. 5. Aufl. Springer-Lehrbuch. Springer Berlin / Heidelberg, 2007.

[FAN11] Yuan Fang et al. „Privacy beyond Single Sensitive Attribute“. In: *Database and Expert Systems Applications*. Hrsg. von Abdelkader Hameurlain et al. Bd. 6860. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, S. 187–201.

[Fed02] Federal Register. *Standard for privacy of individually identifiable health information*. 14. Aug. 2002. URL: http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/index.html (besucht am 15. 03. 2013).

[Fey68] Peter Fey. *Informationstheorie*. 3. Aufl. Berlin, Deutschland: Akademie-Verlag, 1968.

[For08] Otto Forster. *Analysis 1. Differential- und Integralrechnung einer Veränderlichen*. 9. Aufl. Vieweg+Teubner, 2008.

[Fun+10] Benjamin C. M. Fung et al. „Privacy-preserving data publishing: A survey of recent developments". In: *ACM Computing Surveys* 42.4 (Juni 2010), 14:1–14:53.

[Fun+11] Benjamin C. M. Fung et al. „Service-Oriented Architecture for High-Dimensional Private Data Mashup". In: *IEEE Transactions on Services Computing* 99.PrePrints (2011).

[FZ08] Keith B. Frikken et al. „Yet another privacy metric for publishing micro-data". In: *Proceedings of the 7th ACM workshop on Privacy in the electronic society*. WPES '08. Alexandria, Virginia, USA: ACM, 2008, S. 117–122.

[GCG08] Tamas S. Gal et al. „A Privacy Protection Model for Patient Data with Multiple Sensitive Attributes". In: *International Journal of Information Security and Privacy* 2 (3 2008), S. 28–44.

[Gen09] Craig Gentry. „A Fully Homomorphic Encryption Scheme". Dissertation. Stanford University, 2009.

[Gol04] Oded Goldreich. *Foundations of Cryptography: Basic Applications*. 1. Aufl. Bd. 2. Cambridge University Press, Juli 2004.

[Gol06] Philippe Golle. „Revisiting the uniqueness of simple demographics in the US population". In: *Proceedings of the 2006 ACM Workshop on Privacy in the Electronic Society*. (Alexandria, VA, USA, 30. Okt. 2006). Hrsg. von Ari Juels et al. ACM, 2006, S. 77–80.

[GT10] Jacob Goldberger et al. „Efficient Anonymizations with Enhanced Utility". In: *Transactions on Data Privacy* 3.2 (Aug. 2010), S. 149–175.

[Gus97] Dan Gusfield. *Algorithms on Strings, Trees and Sequences. Computer Science and Computational Biology*. Cambridge University Press, 1997.

[Har09] Joachim Hartung. *Statistik. Lehr- und Handbuch der angewandten Statistik*. 15. Aufl. Oldenbourg Verlag München, 2009.

[Her08] Gerd Herold. *Innere Medizin. Eine vorlesungsorientierte Darstellung*. Gerd Herold, 2008.

[HKP12] Jiawei Han et al. *Data Mining : Concepts and Techniques*. 3. Aufl. Bd. 1. The Morgan Kaufmann Series in Data Management Systems. Waltham, MA, USA: Morgan Kaufmann (Elsevier), Juni 2012. URL: http://www.sciencedirect.com/science/article/pii/B9780123814791000010.

[HW97] A. Hundepool et al. „μ-Argus and τ-Argus: Software for Statistical Disclosure Control“. In: *Record Linkage Techniques: 1997 Proceedings of an International Workshop and Exposition*. (Arlington, VA, USA, 20.–21. März 1997). Hrsg. von Wendy Alvey et al. Federal Committee on Statistical Methodology, 1997. Kap. 5, S. 142–149.

[JC05] Wei Jiang et al. „Privacy-Preserving Distributed k-Anonymity“. In: *Data and Applications Security XIX, 19th Annual IFIP WG 11.3 Working Conference on Data and Applications Security*. (Storrs, CT, USA, 7.–10. Aug. 2005). Hrsg. von Sushil Jajodia et al. Bd. 3654. Lecture Notes in Computer Science 19. Springer Berlin / Heidelberg, 2005, S. 166–177.

[JC06] Wei Jiang et al. „A secure distributed framework for achieving k-anonymity“. In: *The VLDB Journal* 15 (Nov. 2006), S. 316–333.

[JX09] Pawel Czeslaw Jurczyk et al. „Distributed Anonymization: Achieving Privacy for Both Data Subjects and Data Providers“. In: *Data and Applications Security XXIII, 23rd Annual IFIP WG 11.3 Working Conference*. (Montreal, QC, Kanada, 12.–15. Juli 2009). Hrsg. von Ehud Gudes et al. Bd. 5645. Lecture Notes in Computer Science 23. Springer Berlin / Heidelberg, 2009, S. 191–207.

[Kan+08] Murat Kantarcioglu et al. „A cryptographic approach to securely share and query genomic sequences“. In: *IEEE Transactions on Information Technology in Biomedicine* 12.5 (Sep. 2008), S. 606–17.

[Kan+09] Murat Kantarcioglu et al. „Formal anonymity models for efficient privacy-preserving joins“. In: *Data & Knowledge Engineering* 68.11 (2009), S. 1206–1223.

[Kar08] Günter Karjoth. „Sind anonymisierte Daten anonym genug?“ In: *digma – Zeitschrift für Datenrecht und Informationssicherheit* 1 (März 2008), S. 18–23.

[KB98] Ronny Kohavi et al. *UCI Machine Learning Repository*. 1998. URL: http://www.ics.uci.edu/%E2%88%BCmlearn/MLRepository.html (besucht am 12. 03. 2013).

[KE11] Alfons Kemper et al. *Datenbanksysteme. Eine Einführung*. 8. Aufl. München, Deutschland: Oldenburg Verlag, 2011, S. 792.

[KJM08] Murat Kantarcioglu et al. „A Privacy-Preserving Framework for Integrating Person-Specific Databases“. In: *Privacy in Statistical Databases, UNESCO Chair in Data Privacy International Conference*. (Istanbul, Türkei, 24.–26. Sep. 2008). Hrsg. von Josep Domingo-Ferrer et al. Bd. 5262. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2008, S. 298–314.

[KL51] S. Kullback et al. „On Information and Sufficiency“. In: *Annals of Mathematical Statistics* 22.1 (1951), S. 79–86.

[Koh05] Wolfgang Kohn. *Statistik. Datenanalyse und Wahrscheinlichkeitsrechnung*. Hrsg. von Holger Dette et al. Bd. 16. Statistik und ihre Anwendungen. Springer Berlin / Heidelberg, 2005.

[KS05] Lea Kissner et al. „Privacy-Preserving Set Operations“. In: *Advances in Cryptology*. CRYPTO 2005: 25th Annual International Cryptology Conference. (Santa Barbara, California, USA, 14.–18. Aug. 2005). Hrsg. von Victor Shoup. Bd. 3621. Lecture Notes in Computer Science 25. Springer Berlin / Heidelberg, 2005, S. 241–257.

[LDR05] Kristen LeFevre et al. „Incognito: efficient full-domain k-anonymity“. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*. (Baltimore, Maryland, USA). Hrsg. von Fatma Özcan. ACM, 2005, S. 49–60.

[LDR06] Kristen LeFevre et al. „Mondrian Multidimensional k-Anonymity“. In: *Proceedings of the 22nd International Conference on Data Engineering*. (Atlanta, Georgia, USA, 3.–7. Apr. 2006). Hrsg. von Karl Aberer et al. IEEE Computer Society, 2006, S. 25–36.

[LLH10] Jinfei Liu et al. „Rating: Privacy Preservation for Multiple Attributes with Different Sensitivity Requirements“. In: *2010 IEEE International Conference on Data Mining Workshops* (2010), S. 666–673.

[LLV07] Ninghui Li et al. „t-Closeness: Privacy Beyond k-Anonymity and ℓ-Diversity“. In: *Proceedings of the 23rd International Conference on Data Engineering*. (Istanbul, Türkei, 15.–20. Apr. 2007). Hrsg. von Rada Chirkova et al. IEEE Computer Society, 2007, S. 106–115.

[LLV10] Ninghui Li et al. „Closeness: A New Privacy Measure for Data Publishing“. In: *IEEE Transactions on Knowledge and Data Engineering* 22 (2010), S. 943–956.

[LS90] D. Linowes et al. „Privacy: the workplace issue of the '90s“. In: *The John Marshall Law Review* 23 (1990), S. 591–620.

[Mac+07] Ashwin Machanavajjhala et al. „ℓ-diversity: Privacy beyond k-anonymity“. In: *ACM Transactions on Knowledge Discovery from Data* 1 (1 März 2007).

[MBM11] Bradley A. Malin et al. „Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule“. In: *Journal of the American Medical Informatics Association* 18.1 (2011), S. 3–10.

[MFD11] Noman Mohammed et al. „Anonymity meets game theory: secure data integration with malicious participants“. In: *The VLDB Journal* 20 (4 2011), S. 567–588.

[Mit12] Hans-Joachim Mittag. *Statistik. Eine interaktive Einführung*. 2. Aufl. Springer-Lehrbuch. Springer Berlin / Heidelberg, 2012.

[Moh+10] Noman Mohammed et al. „Centralized and Distributed Anonymization for High-Dimensional Healthcare Data“. In: *ACM Transactions on Knowledge Discovery from Data* 4.4 (Okt. 2010), 18:1–18:33.

[MW04] Adam Meyerson et al. „On the complexity of optimal K-anonymity“. In: *Proceedings of the Twenty-third ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. (Paris, Frankreich). Hrsg. von Catriel Beeri et al. ACM, 2004, S. 223–228.

[Ner+11] Mehmet Ercan Nergiz et al. „A Look Ahead Approach to Secure Multiparty Protocols“. In: *IEEE Transactions on Knowledge and Data Engineering* 99.PrePrints (2011).

[ØO99] Aleksander Øhrn et al. „Using Boolean reasoning to anonymize databases“. In: *Artificial Intelligence in Medicine* 15.3 (1999), S. 235–254.

[Orl11] Claudio Orlandi. „Is multiparty computation any good in practice?“ In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. (Prag, Tschechische Republik). IEEE Signal Processing Society. IEEE Computer Society, 2011, S. 5848–5851.

[PH78] S. Pohlig et al. „An improved algorithm for computing logarithms over GF(p) and its cryptographic significance“. In: *IEEE Transactions on Information Theory* 24.1 (Jan. 1978), S. 106–110.

[RTG00] Y. Rubner et al. „The earth mover's distance as a metric for image retrieval“. In: *International Journal of Computer Vision* 40.2 (Nov. 2000), S. 99–121.

[RTG98] Yossi Rubner et al. „A Metric for Distributions with Applications to Image Databases“. In: *Proceedings of the Sixth International Conference on Computer Vision Theory and Applications*. (Bombay, Indien, 4.–7. Jan. 1998). IEEE Computer Society, 1998, S. 59–65.

[Sam01] Pierangela Samarati. „Protecting Respondents' Identities in Microdata Release“. In: *IEEE Transactions on Knowledge and Data Engineering* 13 (2001), S. 1010–1027.

[Sch06a] Ingo Schmitt. „Ähnlichkeitssuche in Multimedia-Datenbanken: Retrieval, Suchalgorithmen und Anfragebehandlung“. In: Oldenbourg Wissenschaftsverlag GmbH, 2006. Kap. 5: Distanzfunktionen, S. 165–214.

[Sch06b] Bruce Schneier. *Angewandte Kryptographie . Protokolle, Algorithmen und Sourcecode in C*. Bd. 2. München, Deutschland: Pearson Studium, 2006.

[Sch09] Martin Schumacher. *Methodik Klinischer Studien Methodische Grundlagen Der Planung, Durchfuhrung und Auswertung*. 3. Aufl. Statistik und ihre Anwendungen. Springer Berlin / Heidelberg, 2009, S. 436.

[SH00] Gunter Saake et al. *Datenbanken: Konzepte und Sprachen*. 2. Aufl. mitp Professional. bonn, Deutschland: MITP-Verlag, 2000.

[Sha48] Claude Elwood Shannon. „A mathematical theory of communication“. In: *The Bell System Technical Journal* 27 (Juli 1948). Korrigierter Nachdruck, pages.

[SLZ10] Chaofeng Sha et al. „On *t*-Closeness with KL-Divergence and Semantic Privacy“. In: *Database Systems for Advanced Applications, 15th International Conference*. (Tsukuba, Japan, 1.–4. Apr. 2010). Hrsg. von Hiroyuki Kitagawa et al. Bd. 5982. Lecture Notes in Computer Science 15. Springer Berlin / Heidelberg, 2010, S. 153–167.

[SM88] Egon Seiffart et al. *Lineare Optimierung*. Hrsg. von O. Beyer et al. 4. Aufl. Mathematik für Ingenieure, Naturwissenschaftler, Ökonomen und Landwirte 14. Leipzig: BSB Teubner, 1988. Kap. 2: Die lineare Optimierungsaufgabe, S. 190.

[Spr99] Polly Sprenger. *Sun on Privacy: 'Get Over It'*. Sun Microsystems CEO Scott McNealy über Privacy. Wired News. 26. Jan. 1999. URL: http://www.wired.com/news/politics/0,1283,17538,00.html (besucht am 28. 02. 2013).

[Swe00] Latanya Sweeney. *Simple Demographics Often Identify People Uniquely*. Diskussionspapier zu Data Privacy 3. Pittsburgh, PA, USA: Carnegie Mellon University, 2000. URL: http://dataprivacylab.org/projects/identifiability/index.html.

[Swe01] Latanya Sweeney. „Computational Disclosure Control : A Primer on Data Privacy Protection“. Dissertation. Cambridge, MA, USA: Massachusetts Institute of Technology. Dept. of Electrical Engineering und Computer Science., Juni 2001.

[Swe02a] Latanya Sweeney. „Achieving *k*-Anonymity Privacy Protection Using Generalization and Suppression“. In: *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10.5 (Okt. 2002), S. 571–588.

[Swe02b] Latanya Sweeney. „*k*-Anonymity: a model for protecting privacy“. In: *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10.7 (Mai 2002), S. 557–570.

[Swe03] Latanya Sweeney. *Technologies for Privacy*. Forschungsbericht. School of Computer Science, Carnegie Mellon University, Aug. 2003. URL: http://dataprivacylab.org/center/privacydefs.html#privacyspace (besucht am 12. 12. 2012).

[Swe96] Latanya Sweeney. „Replacing Personally-Identifying Information in Medical Records, the Scrub System“. In: *Proceedings : a conference of the American Medical Informatics Association* (1996), S. 333–337.

[Swe97a] Latanya Sweeney. „Computational Disclosure Control for Medical Microdata: The Datafly System“. In: *Record Linkage Techniques: 1997 Proceedings of an International Workshop and Exposition*. (Arlington, VA, USA, 20.–21. März 1997). Hrsg. von Wendy Alvey et al. Federal Committee on Statistical Methodology, 1997. Kap. 11, S. 442–453.

[Swe97b] Latanya Sweeney. „Guaranteeing anonymity when sharing medical data, the Datafly System“. In: *Proceedings : a conference of the American Medical Informatics Association* (Okt. 1997), S. 51–55.

[TG12] Tamir Tassa et al. „Secure Distributed Computation of Anonymized Views of Shared Databases“. In: *ACM Transactions on Database Systems* 37.2 (2012).

[TV06] Traian Marius Truta et al. „Privacy Protection: *p*-Sensitive *k*-Anonymity Property“. In: *Proceedings of the 22nd International Conference on Data Engineering*. (Atlanta, Georgia, USA, 3.–7. Apr. 2006). Hrsg. von Karl Aberer et al. IEEE Computer Society, 2006.

[Ull88] Jeffrey D. Ullman. *Principles of database and knowledge-base systems*. Bd. 1. New York, NY, USA: Computer Science Press, 1988.

[Ven08] Suresh Venkatasubramanian. „Measures of Anonymity“. In: *Privacy-Preserving Data Mining: Models and Algorithms*. Hrsg. von Charu C. Aggarwal et al. Bd. 34. Advances in Database Systems. Springer US, 2008. Kap. 4, S. 81–103.

[Wei06] Thilo Weichert. „Ein absolutes Muss - Datenschutzmanagement und Technikeinsatz“. In: *Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein* (Mai 2006).

[Wer08] Martin Werner. *Information und Codierung Grundlagen und Anwendungen*. 2. Aufl. Informations- und Kommunikationstechnik. Wiesbaden, Deutschland: Vieweg+Teubner, 2008.

[XT06a] Xiaokui Xiao et al. „Anatomy: Simple and Effective Privacy Preservation“. In: *Proceedings of the 32nd International Conference on Very Large Data Bases*. (Seoul, Korea). Hrsg. von Umeshwar Dayal et al. ACM, 2006, S. 139–150.

[XT06b] Xiaokui Xiao et al. „Personalized privacy preservation“. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*. Hrsg. von Surajit Chaudhuri et al. Chicago, Illinois, USA: ACM, 2006, S. 229–240.

[XT07] Xiaokui Xiao et al. „*m*-invariance: towards privacy preserving re-publication of dynamic datasets“. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*. (Beijing, China). Hrsg. von Chee Yong Chan et al. ACM, 2007, S. 689–700.

[Yao82] A. C. Yao. „Protocols for secure computation“. In: *23rd Annual Symposium on Foundations of Computer Science*. (Chicago, Illinois, USA). IEEE Computer Society, 1982, S. 160–164.

[Ye+09] Yang Ye et al. „Decomposition: Privacy Preservation for Multiple Sensitive Attributes“. In: *Proceedings of the 14th International Conference on Database Systems for Advanced Applications*. DASFAA '09. Brisbane, Australien, 2009, S. 486–490.

[Zha+07] Qing Zhang et al. „Aggregate Query Answering on Anonymized Tables“. In: *Proceedings of the 23rd International Conference on Data Engineering*. (Istanbul, Türkei, 15.–20. Apr. 2007). Hrsg. von Rada Chirkova et al. IEEE Computer Society, 2007, S. 116–125.

[Zho09] Sheng Zhong. „On Distributed k-Anonymization“. In: *Fundamenta Informaticae* 92.4 (Dez. 2009), S. 411–431.

Selbständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe. Weiterhin erkläre ich, eine Diplomarbeit in diesem Studiengebiet erstmalig einzureichen.

Berlin, den 8. Oktober 2013 .....